THE GENOMIC DISTRIBUTION OF HUMAN ENDOGENOUS RETROVIRUSES (HERV-K) AND THE EVOLUTION OF THE PRIMATES

CATRIONA MACFARLANE

A thesis submitted for the Degree of Doctor of Philosophy

The University of Edinburgh

2005

To my family, past and present

I declare that the studies presented here are the result of my own independent investigation. This work has not been submitted for any other degree.

Catriona M. Macfarlane

ACKNOWLEDGEMENTS

I would like to thank my supervisor Prof. Peter Simmonds for all his help, advice and encouragement. I would also like to express my gratitude to all my colleagues in the Virus Evolution Group. I would also like to thank everyone who donated a buccal swab and those who helped in their collection.

My special thanks go to Andrew, my parents Colin and Morag, my sister Heather and my brother Alastair and the rest of my family for all their love, guidance, encouragement and patience.

Contents

Abstract

Endogenous retroviruses (ERVs) are the remnants of ancient germ cell infection by exogenous retroviruses and occupy up to 8 % of the human genome. Following initial infection approximately 28 Million years ago, members of the HERV-K family have continued to amplify and recombine within the genomes of the primate lineage. In this thesis a number of methods have been utilised to investigate the mode of expansion, mechanisms of recombination and the likelihood of HERV replicative activity. Further, HERV polymorphisms were used as phylogenetic markers for the study of human genomic diversity.

Initially a comprehensive catalogue of the total number, cytogenetic location and genomic structure of intact and near intact proviral sequences of HERV-K(HML-2), HERV-K(HML-3) and HERV-K(HML-4) subfamilies was determined within the human genome. As well as highlighting numerous inconsistencies within the HERV literature, six novel proviral sequences were identified.

In this study ERVs were used as phylogenetic markers for primate evolution and speciation for a number of reasons. Firstly, acquisition of an ERV within the germ line represents a unique event in genome evolution and is transmitted as a Mendelian trait in succeeding generations. Secondly, ERVs are homoplasy free traits for which there are no known mechanisms of complete removal without resulting in a telltale deletion of host chromosomal DNA or production of a solitary LTR. Finally, the process of reverse transcription generates two identical long terminal repeats (LTRs) prior to proviral integration; their progressive divergence through

accumulation of nucleotide substitutions through generations of host replication reflects the time since integration.

Comparison of the LTR divergence of each provirus to the physical presence of the element within the primate lineage demonstrated that HERV-K LTRs have been subject to extensive sequence exchange. Thus LTR divergence may not serve as an accurate indicator of time passed since integration. However, the value of ERVs as genetic markers was demonstrated by comparison of their distribution in different human populations which provided novel information on the origin and dispersal of humans over the last 2 million years. Using my catalogue of HERV-K structure and integration site polymorphisms, PCR-based assays were developed for HERV-K loci and used to screen geographically diverse human DNA samples. The results indicate that the diversity of such elements is higher in African than non-African populations with 90.12 % to 99.37 % of genetic variation being within a population. Furthermore, the findings are consistent with an African origin of contemporary humans which was followed by a complex process of interbreeding and population movement.

Investigation of the biological contribution of HERV sequences in serving as nucleation points for chromosomal rearrangement demonstrated that such events have been extremely rare during primate genome evolution and may have been overestimated in previous studies. Analysis of HERV-K coding regions and subfamily phylogeny indicated that they have been subject to both purifying selection and extensive sequence exchange throughout their expansion. Analysis of sequence variation at synonymous and non-synonymous sites in parsimony reconstructed sequences indicates that constraints on sequence variation have reduced over time, suggesting a decline in the likelihood of HERV functionality.

Whilst the role of HERVs in primate evolution is yet to be fully understood, the comprehensive catalogue obtained in this study, the identification of novel proviral sequences and further elucidation of recombinant events provide the foundations for future functional and phylogenetic investigations of human and primate evolution and speciation.

Abbreviations

Alu	A SINE retrotransposon
BLAST	Basic Local Alignment Search Tool
BLASTN	Basic local alignment search tool for nucleotide-nucleotide sequences
BP	Before Present
bp	Base Pair
CA	Capsid
cDNA	Copied DNA
Cytb	Cytochrome b
DNA	Deoxyribonucleic Acid
EDTA	Ethylene-diamine-tetra-acetate
Env	Envelope
ERV	Endogenous Retrovirus
F_{ST}	Wright's Fst statistic, a measure of population differentiation
G^2	Likelihood Ratio
Gag	Group-specific Antigen
HIV	Human Immunodeficiency Virus
He	Heterozygosity
HTDV	Human Teratocarcinoma Derived Virus
HERV	Human Endogenous Retrovirus
HERV-K	Human Endogenous Retrovirus Type K
HML	Human Mouse Mammary Tumour Virus-like
HSPs	High-scoring segment pairs
HTLV	Human T Lymphocyte Virus
IN	Integrase
kb	Kilobase
kDa	Kilodalton
LCR	Low Copy Number Repeat
LINE	Long Interspersed Nuclear Element
LTR	Long Terminal Repeat
Lys	Lysine

MA	Matrix Protein
MHC	Major Histocompatability Complex
MMTV	Mouse Mammary Tumour Virus
mRNA	Messenger RNA
mtDNA	Mitochondrial DNA
Mya	Million years ago
n	Sample size
NC	Nucleocapsid
NCBI	National Centre of Biotechnology Information
ORF	Open Reading Frame
p	Short arm of a chromosome
PBMC	Peripherial Blood Mononuclear Cells
PBS	Phosphate Buffered Saline
PBS	Primer Binding Site
PCR	Polymerase Chain Reaction
PLZF	Promyelocytic Leukemia Zinc Finger Protein
PNG	Papua-New-Guinea
Pol	Polymerase
PPT	Polypurine Tract
PR	Protease
Prt	Protease
q	Long arm of a chromosome
RT	Reverse Transcriptase
SA	Splice acceptor
SD	Splice donor
TM	Transmembrane Protein
RNA	Ribonucleic Acid
rRNA	Ribosomal RNA
tRNA	Transfer RNA
RNAase H	Ribonuclease H
RT-PCR	Reverse Transcriptase PCR
RSV	Rous Sarcoma Virus

SH-IAP	Syrian Hamster Intracisternal-A-Particle
SINE	Short Interspersed Nuclear Element
SU	Surface protein
SVA	SINE.R, VNTR, Alu retrotransposon
UCSC	University California Santa Cruz
UEP	Unique Event Polymorphism
VLP	Virus-like Particle
VNTR	Variable Number Tandem Repeat
v/v	Volume/volume
w/v	Weight/volume
X	X Chromosome
χ^2	Chi-Squared
Y	Y Chromosome

CHAPTER 1

INTRODUCTION

1.1 HERV-K Classification and Discovery

The genomes of humans and other primates contain sequences that resemble retroviruses and are subsequently termed human endogenous retroviruses (HERVs) (Lower et al., 1996). These sequences are vertically transmitted genetic elements that remain from ancient retroviral infection of germ line cells. Following the original insertion of a provirus, intracellular retrotransposition and recombination has led to an increase in the copy number of particular families (Lower et al., 1996). Analysis of the draft sequence of the human genome shows that approximately 8 % is composed of retrovirus-like elements, which includes proviral sequences, derived retrotransposons and a large number of solitary long terminal repeats (LTRs) (Lower et al., 1996; Patience et al., 1997; Lander et al., 2001; Ostertag et al., 2003).

To date, only 26 distinct HERV lineages have been identified, suggesting that all HERVs are derived from only a few germ line invasions by exogenous retroviruses (Tristem, 2000; Benit et al., 2001; Belshaw et al., 2004). HERVs have been divided into three broad classes according to sequence similarities to mammalian exogenous retroviruses. Class I HERVs show similarity to the gammaretroviruses (type C retroviruses); class II to the betaretroviruses and alpharetroviruses (type B and D retroviruses) and class III are related distantly to the spumaretroviruses (Griffiths, 2001). These three major groupings have been further subdivided into families on the basis of sequence similarity and putative tRNA primer binding site specificity (PBS) (Larsson et al., 1989). For example, class II HERVs are often collectively referred to as the HERV-K superfamily as they are primed by a Lysine (Lys) tRNA. However, this nomenclature is ambiguous as highly

divergent HERVs can possess the same PBS tRNA; at least one subgroup is unrelated to its assigned PBS tRNA group (Lavie et al., 2004). These issues are further discussed in Section 3.1.

Phylogenetic analysis has shown that the majority of HERVs entered the genomes of the primate lineage shortly after the divergence of Old and New World monkeys (28 to 45 Mya) although some are also present within New World monkeys (Sverdlov, 2000). The accumulation of stop codons, frameshifts, indels and intra-element recombination leading to the formation of a solitary LTR (Goodchild et al., 1995), has rendered most but not all HERVs transcriptionally inactive (Goodchild et al., 1995; Seifarth et al., 1995; Seifarth et al., 1998; Huh et al., 2003).

Among these, the HERV-K superfamily is acknowledged to be the most biologically active and has been divided into the subgroups HERV-K(HML-1) to HERV-K(HML-10) on the basis of sequence divergence within the *pol* region (Section 3.1). Within this classification, the HERV-K(HML-2) subgroup is of interest as it has retained the ability to encode functional retroviral protein (Towler et al., 1998; Tonjes et al., 1999; Berkhout et al., 1999; Zsiros et al., 1999; de Parseval et al., 2003), produce retrovirus-like particles (Boller et al., 1993; Lower et al., 1993a; Simpson et al., 1996) and includes members which are insertionally polymorphic within humans (Turner et al., 2001; Hughes and Coffin, 2004; Macfarlane and Simmonds, 2004). As this family is central to this thesis, an overview of their history, discovery and classification is presented below.

Members of the HERV-K family were initially discovered by screening human genomic libraries under low stringency conditions using probes related to mouse mammary tumour virus (MMTV) (Callahan et al., 1982; Callahan et al., 1985;

Deen and Sweet, 1986). Twenty five to fifty copies of HERV-K proviruses were then estimated to be present within the human genome using probes designed from the *pol* region of the Syrian hamster intracisternal-A-particle (SH-IAP) provirus (Ono, 1986) and concurrently the prototypic HERV-K10 provirus was also cloned and sequenced (Ono et al., 1986). Two HERV-K proviral forms were identified at this point which were distinguished by a 292 base pair (bp) deletion within the *pol-env* boundary (Ono, 1986). These forms are now designated to the HERV-K(HML-2) subgroup and are referred to as the Type I and Type II proviral genotypes. Type I retain the 292 bp deletion with the prototypic HERV-K10 provirus the standard example.

Cross hybridisation studies using the *gag-pol* MMTV region as a probe revealed the existence of nine different HERV proviral families within the human genome (Franklin et al., 1988). The application of southern blot analysis determined that HERV-K sequences integrated following the evolutionary divergence of New World monkeys but before the split of the Ceropithecoidea and Hominoidea superfamilies (Mariani-Costantini et al., 1989). The expression of HERV-K *gag* and *pol* regions was then observed within peripherial blood mononuclear cells (PBMCs) from healthy individuals (Medstrand et al., 1992) and a full length *gag* open reading frame (ORF) was identified, suggesting that a functional HERV-K protease might exist (Mueller-Lantzsch et al., 1993). Significantly, the HERV-K group was also associated with human teratocarcinoma derived virus particles (HTDV), implying that fully functional HERV-K proviruses could exist within the human genome (Boller et al., 1993; Lower et al., 1993a; Lower et al., 1993b). Sequences associated with these particles were subsequently annotated HERV-K/HTDV. The same year, the HERV-K(HML-1) to HERV-K(HML-6) subgroups were defined on the basis of

sequence diversity within the *pol* region (Medstrand and Blomberg, 1993) and 10,000 to 25,000 HERV-K solitary LTRs were estimated to be present within the human genome (Leib-Mosch et al., 1993).

Detailed phylogentic analysis of the distribution of HERV-K proviruses within the primate lineage, via southern blot analysis using the HERV-K *env* region as a probe, revealed that prosimians and New World monkeys did not contain such sequences within their genomes (Steinhuber et al., 1995) which confirmed the earlier observations of Mariani - Constantini et al., (1989). Notably, within this same year, more than one type of HERV-K subgroup was observed to be expressed within human teratocarcinoma cell lines (Li et al., 1995) and the expression of both Type I and Type II HERV-K(HML-2) genotypes was detected within this same cell line (Lower et al., 1995). Furthermore, a doubly spliced mRNA encoding a protein showing homology to the Lentivirus regulatory protein Rev was identified in human teratocarcinoma cell lines (Lower et al., 1995). Further characterisation revealed the mRNA to be a splice product of the extreme downstream and upstream of the HERV-K(HML-2) Type II *env* region which was subsequently called *cORF* (Lower et al., 1995) and later *Rec* (Magin-Lachmann et al., 2001).

Following these studies, HERV-K elements were associated with virion particles within the human placenta (Simpson et al., 1996) and human mammary teratocarcinoma cell lines (Patience et al., 1996) each of which was also determined to contain a functional reverse transcriptase. Within PBMCs the differential expression of the HERV-K subgroups HML-1 to HML-6 was observed, suggesting a complex system of HERV-K transcription (Andersson et al., 1996). Notably, in the

same year the protease of HERV-K was characterised and confirmed to be fully functional (Schommer et al., 1996).

In 1997, two putative HERV-K proviruses were assigned to chromosomes 7 and 19 and a large number of *gag* and *env* ORFs were mapped within human chromosomes by the application of a protein truncation test in combination with a monochromosomal hybrid mapping panel (Mayer et al., 1997a; Mayer et al., 1997b). Within the same year, HERV-K sequences were linked with type I diabetes (Conrad et al., 1997) however this was later disputed as no association could be found (Lower et al., 1998).

Protein truncation tests were also applied to establish when the 292 bp deletion, which distinguishes HERV-K(HML-2) genotypes, arose within the primate lineage (Mayer et al., 1998). Surprisingly, within older primates a variant of *gag* which was 96 bp longer than HERV-K(HML-2) sequences in the Hominidae family was discovered. This proviral variant was subsequently called HERV-K(OLD) and was verified to be present within the human genome (Reus et al., 2001b).

Proof of the recent retrotranspositional activity of HERV-K was first reported in 1998, when nine HERV-K(HML-2) LTRs were identified as being unique to humans, these LTRs were subsequently ascribed the cluster 9 LTRs (Medstrand and Mager, 1998). The following year, eight human specific HERV-K proviruses which were found (Barbulescu et al., 1999) and a provirus that expressed the proteins *gag* and cORF was assigned to chromosome 7. Interestingly, within the same year, this provirus was shown to carry only a single inactivating point mutation within *pol* which was also observed to be polymorphic in humans (Mayer et al., 1999; Tonjes et al., 1999). It was later estimated that two out of seven individuals will possess the

variant which can encode an intact virus and that this provirus is also polymorphic as a tandemly repeated provirus within contemporary humans (Reus et al., 2001a). Concurrent to the discovery that the HERV-K family has remained active following the evolutionary divergence of humans and chimpanzees, was the identification and characterisation of an active HERV-K reverse transcriptase enzyme (Berkhout et al., 1999) and the observation that HERV-K ORFs appear to be maintained (Zsiros et al., 1999).

Following the initial detection of human specific HERV-K proviruses, a further three proviruses were located which were transcriptionally active (Sugimoto et al., 2001) and several novel HERV-K proviruses were revealed within the human genome via computational screening of the human genome project databases (Costas, 2001; Hughes and Coffin, 2001). A HERV-K provirus was then ascertained to be present within the genomes of chimpanzees and gorillas but not within humans, further emphasising the recent retrotranspositional activity of the family (Barbulescu et al., 2001). Concurrently, two HERV-K proviruses which were insertionally polymorphic within humans were also identified, suggesting that the HERV-K lineage may have been active shortly before the expansion of contemporary human populations (Turner et al., 2001).

The recent retrotranspositional activity of HERV-K(HML-2) was also verified by PCR analysis of solitary LTRs at orthologous regions within non-human primates and subtractive hybridisation (Lebedev et al., 2000; Kurdyukov et al., 2001; Buzdin et al., 2002; Mamedov et al., 2002; Buzdin et al., 2003). Consequently, HERV-K(HML-2) LTRs have been classified into different subgroups on the basis of diagnostic nucleotide differences. This classification is considered in more detail

within Section 3.1. To date, it is estimated that the human genome contains 18 complete and near complete proviruses and 55 LTRs all of which are unique to humans and belong to the HERV-K(HML-2) subgroup (Macfarlane and Simmonds, 2004).

1.2 HERV-K Genome Organisation and Expression

Of the 8 % of the human genome which is composed of retrovirus-like elements the HERV-K(HML-2) subfamily is acknowledged to be the most biologically active (Section 1.1). Consequently, this subfamily will be focused on when reviewing genome organisation and expression. Full length HERV-K(HML-2) proviruses possess ORFs that are recognisable by comparison to the exogenous retrovirus genes *gag*, *prt, pol* and *env* which encode structural proteins, viral enzymes and surface envelope proteins respectively (Figure 1.1). The *gag* gene is located in the 5' upstream region of the genome and the *env* gene the 3' downstream region. The *prt* and *pol* genes are located sequentially between these two genes with *prt* adjacent to *gag* and *pol* adjacent to *env*.

The ORFs have a combined length of approximately 9.2 kilobases (kb) and are flanked by two, non-protein encoding, identical long terminal repeats (LTRs). The LTRs are in turn flanked by 4 to 6 bp chromosomal target site duplications (Figure 1.1). The majority of full length HERV-K(HML-2) proviruses currently recognised within the human genome are designated Type I or Type II, based on the presence (Type I) or absence (Type II) of a 292 bp deletion at the amino-terminal of the *env* gene and carboxy terminus of the *pol* gene (Figure 1.1) (Ono, 1986). Moreover, in contrast to Type II proviruses, the *pol* and *env* genes of Type I sequences are fused (Lower et al., 1995).

In exogenous retroviruses cleavage of the Gag, Prt, Pol and Env precursor proteins form the products of the mature virion (Figure 1.2). By convention these proteins are identified by the following two letter codes; matrix protein (MA), capsid

(CA), nucleocapsid (NC), protease (PR), reverse transcriptase (RT), integrase (IN), surface protein (SU) and transmembrane protein (TM). The structural proteins MA, CA and NC are cleavage products of the Gag precursor protein, PR from Prt, RT and IN from cleavage of Pol and SU and TM from cleavage of Env.

As with other betaretroviruses, such as MMTV, two ribosomal -1 frameshifts are required to translate the 160 kilodalton (kDa) HERV-K Gag-Prt-Pol precursor protein, which is initiated from an AUG start codon located downstream of the 5' LTR (Bannert and Kurth, 2004). Without the ribosomal frameshifts only the 76 kDa Gag protein is translated (Bannert and Kurth, 2004). The first ribosomal frameshift at the 3' end of *gag* and the second at the 3' terminus of *prt* have been shown to occur at a high frequency in MMTV, resulting in an approximate ratio of 1:10 between the frequency of Gag protein expression to the Gag-Pro-Pol polyprotein (Goff, 2001). The Gag polyprotein is autolytically cleaved from the polyprotein by active sites located downstream in the protease domain of Prt (Towler et al., 1998).

In functional retroviruses, such as MMTV, the Gag protein encodes the three major structural proteins MA, NC and CA of mature retroviral particles. CA is the most abundant protein in the virion and forms a core in which two positive stranded genomic RNA molecules are bound to and protected by the NC proteins. MA proteins are located internal to the lipid bilayer (Figure 1.2). Final cleavage of the Gag protein in functional betaretroviruses occurs after particle assembly and budding from the plasma membrane, resulting in condensed core morphology and a mature infectious virion.

Figure 1.1 HERV-K(HML-2) Genome Organisation within the Human Genome.

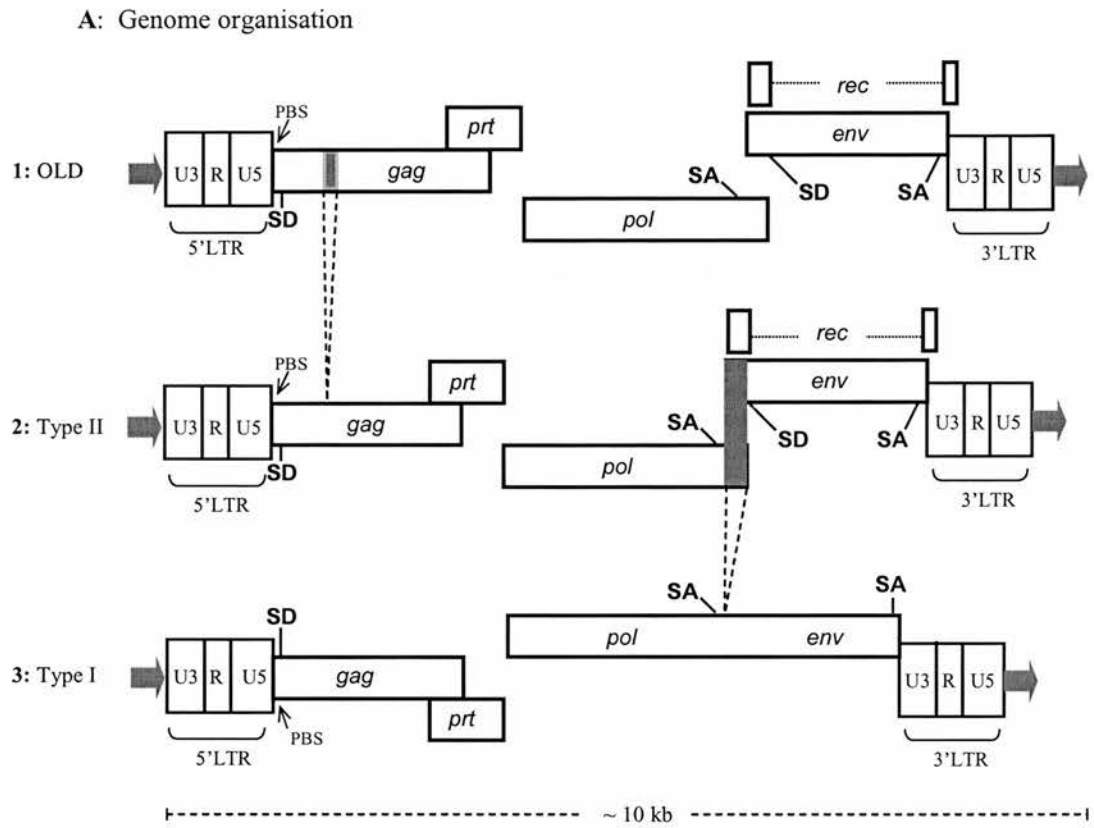
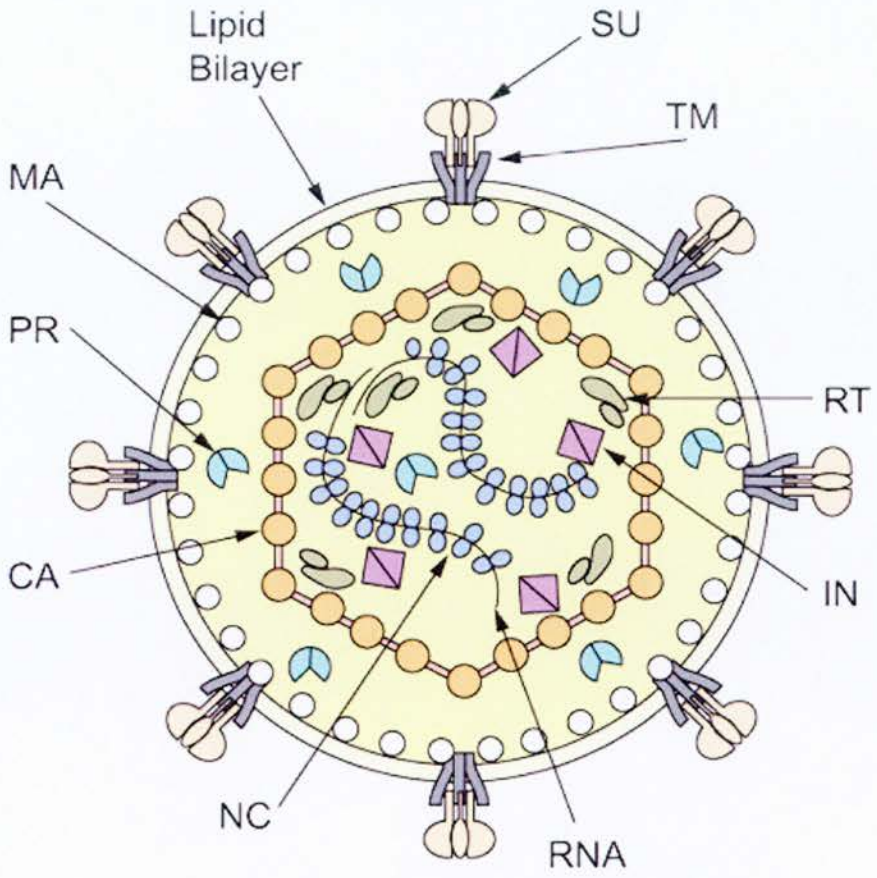


Figure 1.1 HERV-K (HML-2) Genome Organisation within the Human Genome. (**A:** Genome Organisation) HERV-K(OLD) proviral variants are annotated by the number 1, the HERV-K(HML-2) Type II genotypes the number 2 and the HERV-K(HML-2) Type I proviral genomes, the number 3. Grey shading and dashed lines highlight the longer *gag* (96 bp) of HERV-K(OLD) proviruses and the 292 bp deletion within the *pol-env* boundary which distinguishes HERV-K(HML-2) Type I and Type II genotypes. The location of Splice donor (SD) and acceptor (SA) sites are shown. PBS is the tRNA primer binding site and grey arrows flanking each of the LTRs represent the target site duplications generated during viral integration. (**B:** Expected sets of mRNA) Sets of expected mRNAs according to the presence of splice donor and acceptor sites are shown. This figure is a schematic representation adapted from Lower et al., (1995) and Bannert and Kurth, (2004).

Figure 1.2 Schematic cross section through a mature retroviral particle. Viral components are indicated by a two letter code; matrix protein (MA), capsid (CA), nucleocapsid (NC), protease (PR), reverse transcriptase (RT), integrase (IN), surface protein (SU) and transmembrane protein (TM). This figure is reproduced from Voisset and Andrawiss, (2000).



Gag encoding HERV-K(HML-2) transcripts have been detected in a number of healthy and diseased tissue types including teratocarcinoma-derived cell lines, testicular tumour cells and full-term human placentas (Medstrand et al., 1992; Boller et al., 1993; Lower et al., 1993a; Gotzinger et al., 1996; Herbst et al., 1996; Simpson et al., 1996; Herbst et al., 1998; Bieda et al., 2001). For example, in one study, primers designed for HERV-K10 were utilised in an RT-PCR assay of teratocarcinoma-derived cell lines, which subsequently revealed the presence of genome length proviral mRNA and spliced *env* mRNA transcripts (Lower et al., 1993a). In addition, antibodies recognising HERV-K Gag proteins have been observed in a number of additional studies (Sauter et al., 1995; Boller et al., 1997). In one investigation of seminoma and teratocarcinoma patients, 2 % to 4 % of patients exhibited antibodies against recombinant Gag protein compared to 0.1 % to 0.5 % of healthly individuals (Sauter et al., 1995).

The translation of HERV-K *gag* encoded transcripts *in vivo* is confirmed as it has been observed that they form non-infectious virus-like particles (VLPs) in different types of human tissues and cell lines. These include teratocarcinoma-derived cell lines, foetuses and both malignant and non-malignant breast samples (Mondal and Hofschneider, 1982; Al Sumidaie et al., 1988; Lower et al., 1993a; Bieda et al., 2001). In addition, recombinant HERV-K10 *gag* regions from cellular DNA have been used to raise antibodies against Gag protein and were shown to stain VLP particles in immunoelectron microscope studies with human teratocarcinoma-derived cell lines (Lower et al., 1993a). Antibody responses to HERV-K(HML-2) encoded proteins provide compelling evidence of proviral gene expression. However, in a number of studies the exact antigen that elicited an antibody response or the

epitopes involved remain unclear. As such, cross reactivity with unrelated proviral, viral or cellular proteins could potentialy have elicited the response (Bannert and Kurth, 2004).

It has been suggested that VLPs may represent pseudotypes generated by complementation of several expressed protein products in *trans* from HERVs which are defective in *cis* (Seifarth et al., 1995). Moreover, the analysis of VLPs released by human mammary carcinoma derived cell lines, which exhibited type B morphology, showed that the particles packaged retroviral genomes which were defective and encoded by varying proviral sequences (Seifarth et al., 1995). One of the three proviral sequences was later determined to belong to the HERV-K(HML-4) subgroup (Seifarth et al., 1998). These results indicate that in addition to potential complementation in *trans* during replication, genomes encoded by highly divergent proviruses can be packaged within the same virion. Furthermore, VLPs package defective proviral genomes which are unrelated to the provirus which encoded the Gag structural proteins required for particle formation (Seifarth et al., 1995; Seifarth et al., 1998).

Retroviruses require the Gag and Env proteins to co-localise with the Gag-Prt-Pol polyprotein to form replication competent virion particles. Packaging of RNA into viral particles during co-localisation is mediated by specific interactions between viral RNA, packing signals and the NC protein of Gag. In betaretrovirus, proteolytic cleavage of the polyprotein by PR occurs after budding from the cellular phospholipid bilayer, resulting in the release of the structural proteins of Gag, PR of Prt, and RT and IN of Pol.

The observed lack of mature condensed core regions and an electron lucent space between the core and the viral membrane suggest arrested VLP maturation, which is consistent with a failure of Gag protein cleavage (Tonjes et al., 1997). This could be due to Gag proteins failing to undergo correct downstream processing or a lack appropriate recognition signals which prevent cleavage by the virally encoded PR (Bannert and Kurth, 2004). In addition, arrested maturation of VLPs is further implied by a lack of envelope spikes and the observation particles are very rarely seen separated from the plasma membrane (Boller et al., 1993; Tonjes et al., 1997; Bieda et al., 2001). It has been suggested that deficiencies in VLP morphogenesis may be due to a transdominant effect, in which incorporated mutated proteins are encoded by multiple proviruses (transcomplementation), this may additionally interfere with viable particle formation (Lower et al., 1996).

The *prt* gene of full length HERV-K(HML-2) Type I and Type II proviruses encodes a putative aspartic PR protein which exhibits close homology to that of other betaretroviruses (Ono et al., 1986). Recombinant HERV-K10 PR has been generated and shown to possess a fully functional protease core which can autolytically cleave to produce a smaller 18 kDa fragment, which also retains proteolytic activity (Schommer et al., 1996; Towler et al., 1998). Interestingly, it has been demonstrated that recombinant HERV-K10 PR is capable of cleaving the Human Immunodeficiency Virus (HIV) Gag protein at authentic HIV-PR recognition sites (Towler et al., 1998).

The *pol* gene of replicative competent retroviruses encodes both RT and IN that catalyse the reverse transcription of the genomic RNA and integration of the subsequent cDNA molecule into the host genome. Pol is proteolytically cleaved by

virally encoded PR during particle maturation. The *pol* gene of HERV-K10 encodes a protein analogous to the MMTV Pol (Ono et al., 1986) and a number of HERV-K(HML-2) proviruses have been shown to possess *pol* ORFs capable of encoding functional RT and IN enzymes (Barbulescu et al., 1999; Turner et al., 2001). For example, six HERV-K10 proviral RT encoding regions were cloned and used to generate recombinant proteins, five of which exhibited polymerase activity (Berkhout et al., 1999). The recombinant RT also demonstrated ribonuclease H (RNase H) activity that is intrinsic to retroviral RT enzymes and degrades the RNA template after it has been reverse transcribed (Berkhout et al., 1999). In an additional study, the putative IN encoding ORF from the HERV-K10 *pol* gene, which included two motifs common to retroviral IN, was cloned and used to express an IN fusion protein (Kitamura et al., 1996). This fusion protein was highly active for both terminal cleavage and strand transfer for HERV-K10 LTR substrates and was also active against both HIV and Rous Sarcoma Virus (RSV) LTRs. Conversely, IN encoded by HIV and RSV was not catalytic against HERV-K10 LTRs (Kitamura et al., 1996).

Viral particles containing or associated with active RT and PR have been isolated from a number of human tissues including normal placenta, platelets, from patients with thrombocythemia, cell culture supernatant derived from diabetic pancreases, breast cancer and teratocarcinoma derived cell lines (Seifarth et al., 1995; Patience et al., 1996; Simpson et al., 1996; Boyd et al., 1997; Berkhout et al., 1999). Studies have also shown that particles associated with RT and PR activity, package HERV-K(HML-2)-like RNA sequences (Seifarth et al., 1995; Patience et al., 1996).

The *pol* and *env* reading frames of the Type II proviral variants overlap and so require the production of a spliced transcript prior to translation of Env (Figure 1.1). Due to the 292 bp deletion at the *pol-env* boundary of Type I sequences, such proviruses are unable to produce complete spliced transcripts for the translation of either Pol or Env polypeptides (Ono et al., 1986).

In active exogenous retroviruses *env* encodes a glycoprotein which is localised to the lipid bilayer of the retroviral particles. In such viruses, Env glycoproteins are cleaved into two polypeptides by the viral protease resulting in a surface subunit (SU), which is exposed on the virion surface, and a single-pass transmembrane protein (TM) anchored to the matrix protein (Figure 1.2). Both Env glycoproteins are required for cell surface receptor recognition and mediation of fusion between viral and host cell lipid bilayer membranes. Of the two, it is believed that the SU subunit is the major factor in cellular recognition and that TM plays the most important role in lipid bilayer fusion.

Spliced *env* transcripts have been detected in a number of cell types and tissues including placenta, testis, testicular tumours, melanomas, prostate and ovarian melanomas (Lower et al., 1993a; de Parseval et al., 2003). Antibodies to HERV-K Env proteins are frequently detected in patients with germ-cell tumours (Boller et al., 1997) and expression of Env protein has been demonstrated in many human breast cancers and at higher levels in patients with HIV (Wang-Johanning et al., 2001; Wang-Johanning et al., 2003).

An alternative doubly spliced *env* transcript, called cORF, has been demonstrated within HERV-K(HML-2) Type II proviruses and encodes the protein Rec (Lower et al., 1993a). Rec is a 14 kDa protein which shares its 87 upstream

amino acids with the Env protein. The downstream 18 amino acid residues of Rec are encoded by a region directly upstream and overlapping into the 3'LTR, which is in an alternative reading frame to *env* (Figure 1.1). Rec shows a high degree of functional homology to the RNA-binding nuclear export proteins of HIV (Rev) and human T lymphotropic Virus (HTLV) (Rex) (Lower et al., 1995). It also appears to be the only regulatory protein potentially encoded by HERV-K(HML-2) (Magin-Lachmann et al., 2001). The splice acceptor (SA) and splice donor (SD) sites required for the generation of the cORF transcript are situated within the *pol-env* boundary, consequently HERV-K(HML-2) Type I proviral genotypes are unable to produce transcripts encoding the Rec protein (Lower et al., 1995).

As with the HIV encoded Rev, Rec has been shown to bind and stabilise partially or completely spliced viral transcripts and enhance their nuclear export (Boese et al., 2000b; Magin-Lachmann et al., 2001). The Rec binding site is a region of highly structured RNA in the U3 domain of the 3'LTR, which has also been demonstrated to bind Rev and Rex from HIV and HTLV respectively (Magin-Lachmann et al., 2001). However, Rec is not able to substitute for Rev in HIV (Magin-Lachmann et al., 2001). It has also been shown that Rec interacts with the host cellular factor Promyelocytic Leukaemia Zinc Finger Protein (PLZF) (Boese et al., 2000a; Boese et al., 2001). PLZF plays a role in the regulation of cell growth, differentiation and apoptosis. Expression of HERV-K(HML-2) Rec protein has been demonstrated to induce tumour formation in nude mice and is associated with germ cell tumours such as testicular cancer (Boese et al., 2000a).

The HERV-K(HML-2) Type I proviral variants are unable to produce Rec transcripts due to their lack of splice acceptor (SA) and splice donor (SD) sites.

However, a smaller 9 kDa protein called Np9 has been associated with HERV-K(HML-2) Type I (Armbruester et al., 2002). This protein is encoded from transcripts generated using an alternative splice-donor site directly upstream of the 292 bp deletion (Armbruester et al., 2002) (Figure 1.1). Np9 shares only the 15 N-terminal amino residues with Rec with the remainder of the protein encoded from a third reading frame which is not shared by *env* or *rec*. An increased predominance of Np9 transcripts has been detected in human mammary carcinomas (Armbruester et al., 2002).

Proviral transcription is controlled by promoters, enhancers, polyadenylation signals and various transcriptional regulators located within the LTRs. HERV-K(HML-2) LTRs are approximately 970 bp in length and contain three distinct domains U3, R and U5 (Figure 1.1). The U5 and R regions contain regulatory elements, including negative regulatory domains. In the HERV-K(HML-2) proviral variants, the U5 region of the 5'LTR overlaps with the *gag* ORF and the U3 region of the 3'LTR overlaps with the *env* ORF (Figure 1.1).

HERV-K(HML-2) LTRs have been shown retain enhancer and promoter activity for the expression of a luciferase reporter gene in a number of human cell lines (Ruda et al., 2004). However, the promoter activity was observed to vary between cell types. An earlier study indicated that the ability of HERV-K(HML-2) LTRs to promote and enhance expression of a luciferase gene is bi-directional, relative to the reporter gene and demonstrated the presence of a silencer like element in the U5 region (Domansky et al., 2000). It has also been shown that the 5' region of the HERV-K(HML-2) U3 domain specifically binds host nuclear proteins in a tissue

specific manner and that this may play a role in the observed tissue specific promoter activity (Akopov et al., 1998).

1.3 Retroviral Life Cycle

Infection with an extracellular retrovirus is initiated by the binding of the SU envelope glycoprotein to a specific receptor on the host cell surface resulting in fusion of the virus envelope with the cellular membrane and entry into the host cell via-receptor-mediated endocytosis (Figure 1.3). Entry into the cell is associated with partial uncoating and release of the core particle (nucleocapsid) into the cytoplasm. Reverse transcription of the retroviral genomic RNA sequence occurs within the core complex which includes NC, RT, IN and viral RNA. The signal that triggers reverse transcription is unclear; although it has been speculated that exposure of the core to high levels of deoxyribonucleotides is sufficient (Goff, 2001).

Reverse transcription utilises virally encoded RT and is primed at the PBS by cellular tRNA, both of which are packaged within the infecting retroviruses core structure. Reverse transcription is initiated from the paired 3' OH of the tRNA, which anneals to a complementary PBS sequence located downstream of the U5 region of the infecting viral genome (Figure 1.4). Elongation from the tRNA primer produces a U5 and R complementary cDNA intermediate, termed a minus-strand strong-stop DNA and results in the degradation of the transcribed template RNA through the RNase H activity of RT. Degradation of the complimentary RNA exposes the single stranded minus-strand cDNA molecule facilitating the translocation and annealing of the strong-stop intermediate to the complimentary R sequence at the 3' of the RNA template. Sequence analysis of HIV progeny LTRs indicates that retrovirus strand transfer may occur in both *cis* and *trans* between the two packaged RNA genomes (Yu et al., 1998). Strand elongation of the

Figure 1.3 Retroviral Life Cycle. Adapted from Sverdlov, (2000). Retroviral encoded proteins are represented by yellow circles (RT), orange squares (IN) and blue squares (structural proteins).

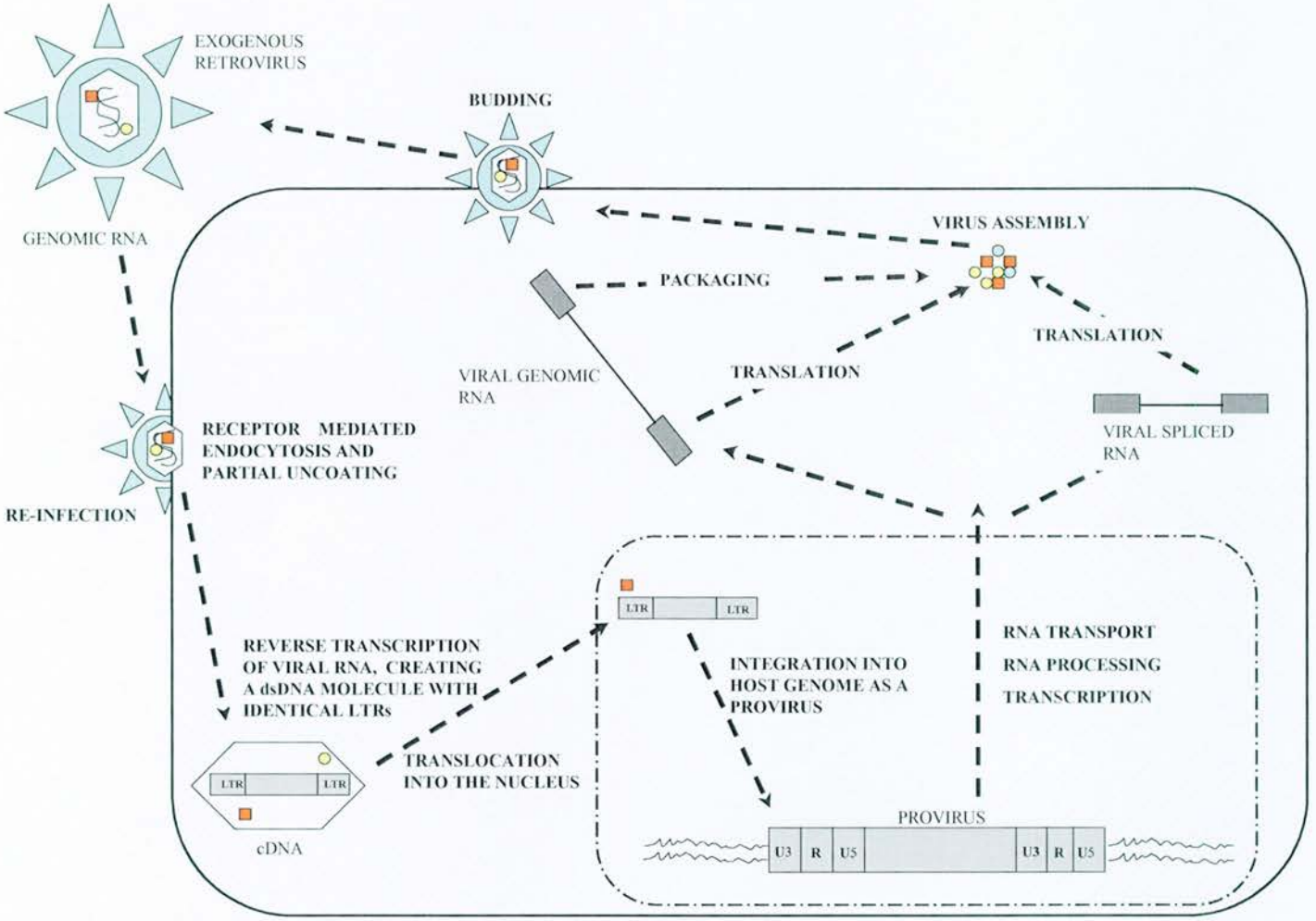
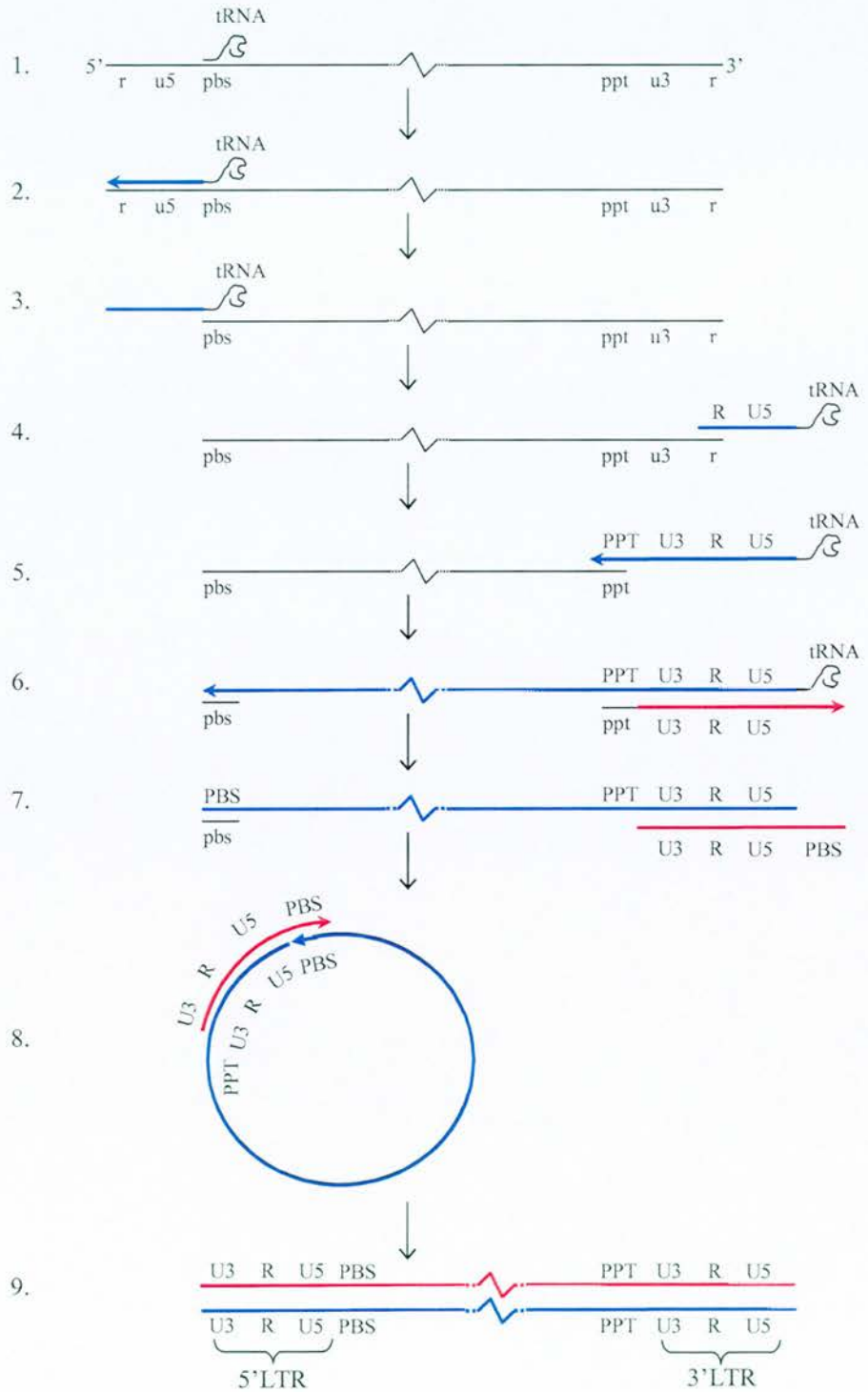


Figure 1.4 Reverse Transcription of the Retroviral Genome. Thin lines represent viral RNA, thick lines represent cDNA and shaded arrows represent progress and direction of transcription. The wavy line indicates lack of scale. Adapted from Goff, (2001).



minus-strand cDNA by RT is primed from the annealed strong-stop DNA with strand elongation forming a complementary cDNA product and degrading the majority of the positive stranded RNA template (Figure 1.4).

Elongation of the positive stranded DNA is initiated from a purine rich template RNA sequence in the polypurine tract (PPT) domain which is relatively resistant to the RNase H activity of RT. Extension of the positive stranded cDNA proceeds downstream copying the U3, R and U5 regions until it reaches the PBS complementary sequence of the still bound tRNA. The tRNA contains a modified base which results in the termination of transcription.

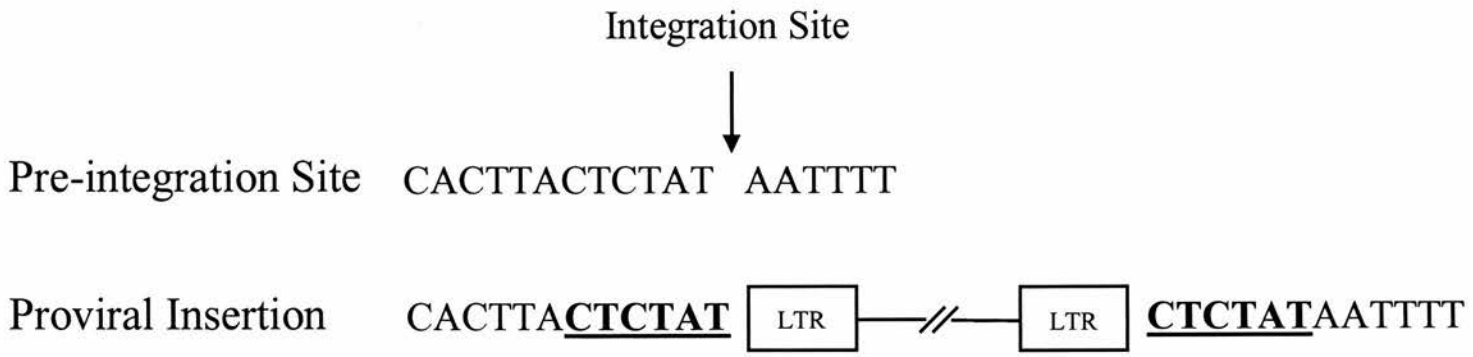
The tRNA is then removed by the RNase H activity which is predominant at the cDNA-RNA boundary. Removal of the tRNA permits the annealing of the positive and negative sense complementary PBS sequences, during a second translocation event. This results in the formation of a circular cDNA intermediate. A final elongation step extends both the positive and negative strands to completion, which also displaces the paired PBS sequences. The final result is a linear double stranded cDNA molecule with identical inverted 5' and 3' LTRs that contain sequentially located U3, R and U5 domains.

The double stranded cDNA migrates to the nucleus and is integrated into the host genome forming a provirus (Figures 1.3 and 1.4). In most simple retroviruses entry to the nuclease is dependent on the breakdown of the nuclear membrane during cell division (Goff, 2001). Integration of the retroviral cDNA into the host genome involves three major steps and is mediated by IN packaged within the infecting viral core structure. Initially, the endonuclease activity of IN catalyses the removal of a dinucleotide from a highly conserved CA motif at the 3' ends of both the positive

and negative strands of the retroviral cDNA molecule. This is followed by a strand transfer reaction in which the host target DNA is nicked and phosphdiester bonds form between nicked target DNA phosphate backbone and the 3' OH ends of the retroviral cDNA, created by the endonuclease activity of IN. The nicking process is staggered by a small number of base pairs (generally between 4 to 6 bps) which results in strand transfer and the reproduction of the site of integration within the host. Such regions are called target site duplications and are located directly upstream and downstream of the provirus (Figure 1.5). Finally, the single stranded target site duplications are filled and joined to the unpaired 5' proviral nucleotides by the host cellular repair machinery. However, it has also been speculated that IN may exhibit a polymerase activity which plays a role in filling the gaps left by the target site duplications (Goff, 2001).

Transcription of the provirus is dependent on host cellular RNA polymerase II machinery. Transcribed proviral RNA is either packaged into new retroviral particles, translated by cellular machinery in spliced and unspliced genomic forms, or the RNA may reintegrate in the host genome (Figure 1.3). In the final stages of replication viral proteins are assembled in the cytoplasm and packaged with retroviral RNA before budding from the cell membrane. During the budding process the viral particles acquire a lipid envelope and the virally encoded envelope glycoproteins. The final maturation process of many retroviruses, such as betaretrovirus, involves the cleavage of the Gag, Gag-Prt-Pol and Env polyproteins by the protease domain of PR, resulting in infectious viral progeny which exhibit a distinctive morphology.

Figure 1.5 Host Genome Sequence before and after the Integration of a Retrovirus. The nucleotides highlighted in bold and underlined represent the target site duplications generated upon entry of the retroviral sequence into the host genome.



Many HERV-K(HML-2) sequences posses ORFs capable of encoding the Gag and Gag-Prt-Pol polyproteins (Sections 1.1 and 1.2). A number of studies have demonstrated active PR, RT and IN within host cells and associated with extracellular VLPs (Section 1.2). The expression of functional HERV-K(HML-2) viral enzymes and the possession of intact LTRs within several proviruses suggests that this subgroup may be capable of a number replication steps associated with exogenous retroviruses. These include reverse transcription and integration of transcribed proviral RNA into the host genome.

HERV-K(HML-2) encoded Gag structural proteins have been shown to be present within VLPs which were at the stage of budding from the cellular membrane (Mondal and Hofschneider, 1982; Al Sumidaie et al., 1988; Lower et al., 1993a; Bieda et al., 2001). These observations suggest that a number of HERV-K(HML-2) encoded proteins retain the ability to undergo a process resembling virus assembly, RNA packaging and budding. However, the morphology of the VLPs implies that the Gag and Env proteins are unable to undergo the final maturation step of cleavage to produce mature structural proteins and envelope glycoproteins (Section 1.2) (Boller et al., 1993; Tonjes et al., 1997; Bieda et al., 2001). The lack of mature surface glycoprotein spikes would prevent receptor recognition and receptor-mediated endocytosis. Furthermore, transcomplementation of viral proteins encoded by different HERV families may inhibit VLP infectivity (Lower et al., 1996).

Consequently, based upon the VLP morphology and the lack of evidence for a life cycle wholly comparable to an exogenous retrovirus (Lower et al., 1996; Bannert and Kurth, 2004), it is highly unlikely that VLPs encoded by HERVs

proliferate via reinfection and are probably restricted to a mechanism of intracellular retrotransposition.

1.4 Biological Impact of HERV-K upon the Primate Lineage

Following the insertion of a HERV into host chromosomal DNA, the proviral sequence remains stable as it cannot be easily removed. Mechanisms of removal result in either the deletion of flanking DNA, which is likely to affect host phenotype, or the loss of proviral internal genic regions which leaves behind a solitary LTR (Macfarlane and Simmonds, 2004). The long term survival of a retrovirus-like sequence and the continued amplification of HERV families over long evolutionary timescales have fuelled speculation of the biological contribution of HERVs to the primate lineage. The outcome of continued HERV proliferation within the genomes of the primate lineage may be expected to have had both pathogenic and non-pathogenic effects.

As with any other retroelement, the retrotranposition and insertion of an HERV into the host genome can generate genetic instability (Deininger and Batzer, 2002). To date, there are no reported inherited human diseases associated with the *de novo* insertion of an HERV-K provirus. However, the number of HERV-K integration events that are specific to humans is significantly less than those reported for other retroelement families such as LINE and Alu (Deininger et al., 2003; Macfarlane and Simmonds, 2004). Interestingly, the retrotransposon family SVA, which is derived from HERV-K(HML-2) sequences has been shown to be actively retrotransposing within contemporary humans (Bennett et al., 2004) with one novel insertion leading to hereditary elliptocytosis (Ostertag et al., 2003).

The abundance of HERV sequences within a genome can also lead to genomic instability whereby they serve as nucleation points for recombination

events. It has been estimated that at least 16 % of HERV-K(HML-2) proviral sequences contained within the human genome have been involved in non-allelic recombination events, which are likely to have had a major impact on primate genome evolution (Hughes and Coffin, 2001). However, with the exception of the non-allelic recombination between two HERV-15 proviruses on the Y chromosome (Blanco et al., 2000; Kamp et al., 2000; Sun et al., 2000; Bosch and Jobling, 2003), human diseases have not been associated with recombination between HERV sequences. The role that HERVs have played in shaping primate genomes is considered in more detail in Chapter 5.

A further proposed impact of HERV-K upon the primate lineage is the regulation of gene expression (Sverdlov, 2000). As HERV-K LTRs contain putative hormone responsive elements, enhancers, promoters, silencers, polyadenylation signals and transcription factor binding sites they could cause significant changes in the regulation of neighbouring genes (Lavrentieva et al., 1998). Examination of the physical locality of HERV-K LTRs has indicated that they tend to be situated in transcriptionally active regions (Leib-Mosch and Seifarth, 1995) with some in close vicinity to genes (Andersson et al., 1998; Lavrentieva et al., 1998; Liao et al., 1998; Kurdyukov et al., 2001; Buzdin et al., 2003). Interestingly, analysis of HERV-K LTRs on chromosome 19 has shown that they tend to be located close to zinc finger genes (Lavrentieva et al., 1998) and HERV-K LTRs have been shown to have retained the ability to bind cellular protein factors (Akopov et al., 1998). HERV-K(HML-2) LTRs are observed to possess bi-directional promoter activity (Domansky et al., 2000) whereas HERV-K(HML-4) LTRs have only been observed to be active in a forward direction (Seifarth et al., 1998; Baust et al., 2001). The

promoter activity of HERV-K LTRs varies between cell types (Vinogradova et al., 2001; Ruda et al., 2004), implying that they may be involved in specific regulatory pathways. Involvement of ERVs in influencing gene expression within humans and other species is highly documented (Brosius, 1999) however, only a few examples are attributed to the HERV-K superfamily. A summary of these is presented below (Table 1.1).

The expression of HERV-K proteins could also have a major impact upon the primate genome either by conferring resistance to retroviral infection (Sverdlov, 2000) or leading to disease. To date, the causative or disease-promoting association of HERV-K protein expression within germ cell tumours and mammary carcinomas has yet to be confirmed (Bannert and Kurth, 2004). However, there is strong evidence for the involvement of HERV-K(HML-2) in malignancy (Section 1.2). For example, the expression of the HERV-K(HML-2) Rec protein has been demonstrated to induce tumour formation in nude mice and is associated with germ cell tumours such as testicular cancer (Boese et al., 2000a). The HERV-K(HML-2) subgroup has also been implicated in the autoimmune disease insulin-dependent diabetes mellitus (IDMM). A superantigen domain within the *env* region of HERV-K18 was detected in the pancreatic islets of diabetic patients (Conrad et al., 1997), however, this was later refuted as no association could be found (Lower et al., 1998; Kim et al., 1999).

It should be noted that the genomic retroviral elements that exist today represent only a small fraction of total germ line integration events, namely those that were not detrimental to the host and that became fixed within the genomes of the primate lineage.

Table 1.1 HERV-K LTRs which are involved in Gene Expression. Adapted from Sverdlov, (2005)

HERV	Human Gene	Function	Reference
HERV-K(HML-2)	*LEPR* (leptin receptor)	Polyadenylation	(Kapitonov and Jurka, 1999)
HERV-K(HML-4)	*FLT 4* (transmembrane tyrosine kinase)	Polyadenylation	(Baust et al., 2000)
HERV-K(HML-6)	*HLA-DRB6* (histocompatibility complex)	Promoter	(Mayer et al., 1993)
HERV-K(HML-8)	*LEP* (leptin)	Enhancer	(Bi et al., 1997)
HERV-K(HML-10)	*INSL4* (early placental insulin-like peptide)	Promoter	(Bieche et al., 2003)

1.5 Primate Evolution

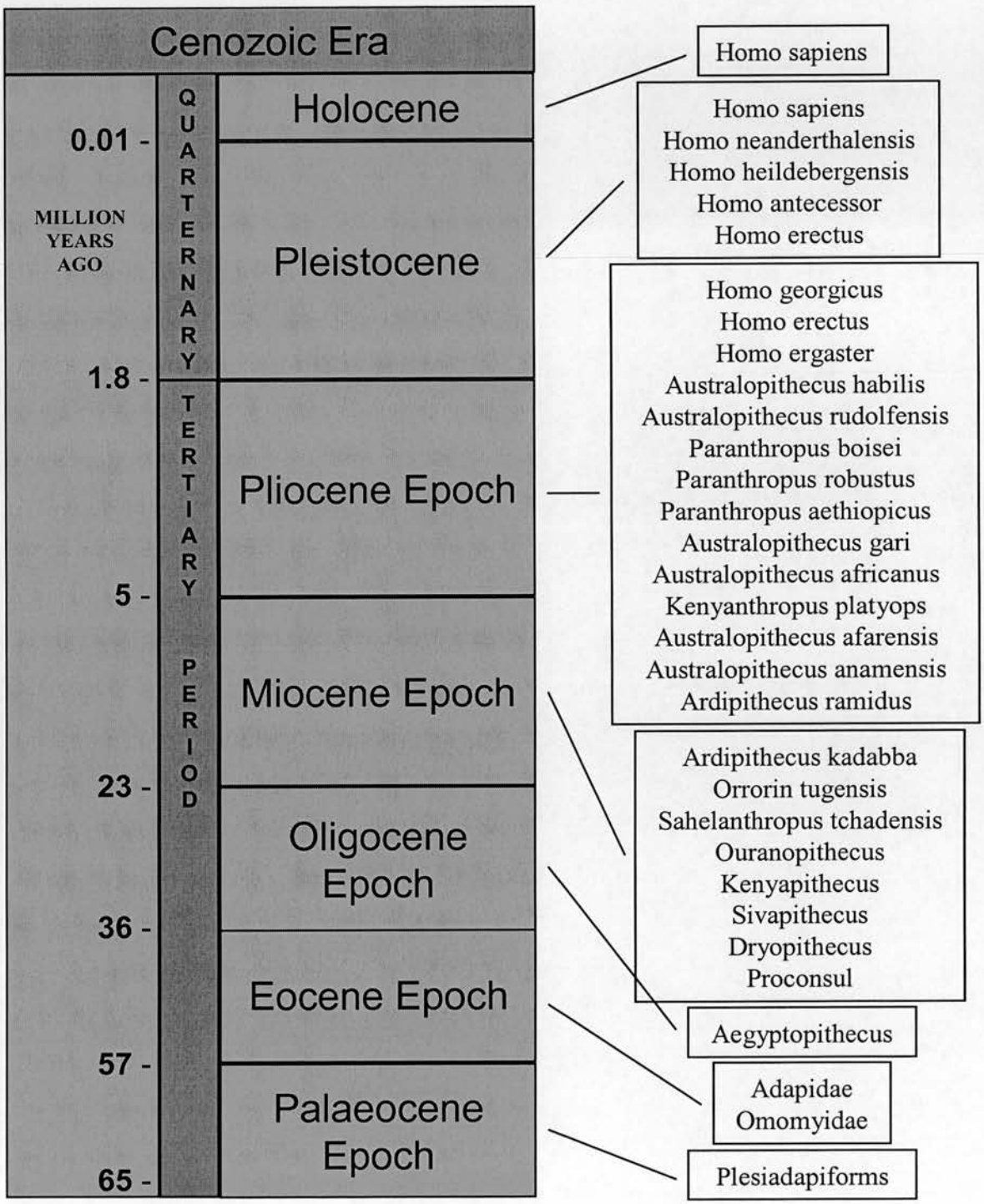
1.5.1 Evolution of the Primates

The Cenozoic is the Earth's current geological era which begins at approximately 65 Mya (Figure 1.6). The beginning of the era is defined by global mass extinction which resulted in loss of 50 % of extant genera including the dinosaurs and is attributed to a combination of volcanic eruptions (Keller et al., 2004), climatic cooling and a meteorite collision which generated the Chicxulub impact crater off the Yucatan Peninsula (Pope et al., 1998).

The first mammals appear approximately 200 Mya during the Mesozoic era and by the Cenozoic were highly diversified in diet, morphology and behaviour to facilitate their survival of the mass extinction; as the large bodied vertebrates died, mammals filled ecological niches by adaptive radiation (Hedges et al., 1996). The first archaic primates *Plesiadapiformes* appear at the beginning of the Palaeocene Epoch approximately 65 Mya (Figure 1.6) and by 60 Mya were widespread in the tropical climates of the northern hemisphere. At this time India and Africa were islands, North America was connected to Europe and Asia forming a northern continent called Laurasia and South America and Australia were connected to Antarctica forming a southern continent called Gondwana.

By the middle Palaeocene (60 Mya) four families of *Plesiadapiformes* were present; *Pleisadapsis* made up the largest family and is presumed to be ancestral to the later *Adapidae* which is considered be one of the first euprimates (primates of modern aspect). At the end of the Palaeocene global temperatures increased leading

Figure 1.6 Cenozoic Timescale and Associated Hominids, Hominoids and Proto-Hominoids



to the replacement of tropical vegetation with dense sub-tropical forests; this resulted in a decline of the *plesiadapiformes* with only one species, *Paromomyidae*, surviving in North America at 35 Mya.

The Eocene (36-57 Mya) is characterised as being the warmest epoch during the tertiary period with vegetation around the globe being predominantly sub-tropical, this facilitated movement and diversification of the euprimates across the Laurasian continent. During this epoch North America separated from Eurasia, South America migrated north away from Antarctica and India collided with the Eurasian continent generating the Himalayas and the Alps. The beginning of the Eocene epoch also corresponds with the first appearance of the Prosimians (Figure 1.7), they are classified into the *Adapidae* which are believed to be the ancestors of present day Lemurs and Lorises and *Omomyidae* which gave rise to Tarsiers (Seiffert et al., 2003; Bloch and Silcox, 2001). Controversy surrounds which of these lineages later gave rise to apes and humans, the consensus is that *Adapidae* gave rise to the *Anthropoidea* (anthropoids) as fossilised examples of the *Omomyidae* show highly specialised auditory morphology (Beard et al., 1991).

By the middle Eocene (50-55 Mya) the euprimates were abundant on several continents but were absent from South America and Antarctica. They had a highly developed morphology which included grasping hands with nails and reliance upon sight rather than smell. The primate family *Eosimiidae* is intermediate between the anthropoids and prosimians and as examples are present in China at 42 Mya, this would suggest that the anthropoid primates initially arose within the sub-tropical regions of Asia (Gebo et al., 2000; Jaeger et al., 1999). From the middle Eocene global temperatures began to cool which denoted the beginning of a 10 Mya

Figure 1.7 Classification of the Primates

Order	Suborder	Infraorder	Superfamily	Family	Subfamily	Tribe	Groups
PRIMATES	Prosimians						Loris Lemurs Tarsiers
	Anthropoidea	Platyrrhini					New World Monkeys
		Catarrhini	Cercopithecoidea				Old World Monkeys
			Hominoidea	Hylobatidae			Gibbons
				Hominidae	Ponginae		Orang
					Homininae	Panini	Gorillas Chimps
						Hominini	Humans

fluctuation in climate and environment which also defines the boundary between the Eocene and Oligocene epochs.

In addition to the lowering of sea levels and the formation of an ice-cap over Antarctica, one of the environmental consequences of global cooling at the beginning of the Oligocene (36 Mya) was the retreat of sub-tropical forests to regions along the equator; consequently, primate species became isolated within Africa, Arabia and Asia. The Fayum deposits of Egypt are the most well studied sub-tropical region from this time and at the beginning of the Oligocene (33-35 Mya) they contain several significant anthropoid primates including the Catarrhine primates, *Agyptopithecus* and *Propliopithecus.* The morphology of *Agyptopithecus* is noteworthy as it appears to be an intermediary between the Old World monkeys and later Miocene Apes as it possesses post-cranial and cranial characteristics of the monkeys but a dentition pattern more closely related to the Apes (Fleagle and Simons, 1982).

The Oligocene is also embodied by the division of the infraorder of monkeys with New World monkeys (Platyrrhine) appearing in South America approximately 30 Mya (Takai et al., 2000). It is unclear how geographically this subdivision occurred as South America was an island continent for much of the Teritary period with separation from Gondwana beginning at approximately 55 Mya and connection to North America occurring between 2 and 3 Mya. Several theories have been proposed, including two which require the crossing a strait of water by a combination of island hopping or rafting, with origins from either North America or Africa. An African origin is favoured as a North American origin of the Platyrrhine (anthropoid) monkeys would require the parallel origin and evolution of the prosimians

(Strepsirrhine) to the anthropoid form on both the North American and African continents. To further resolve this anomaly, several studies have examined molecular phylogenetic relationships between extant primates and attempted to define the time of divergence of each of the genera (Table 1.2). The conclusion of these and other morphological studies is that the South American monkeys most closely resemble the African monkeys with division approximately 35 Mya (Goodman et al., 1994; van der Kuyl et al., 1995; Shoshani et al., 1996; Noda et al., 2001).

The Miocene (5-23 Mya) epoch begins with the first appearance of the first true apes or tailless primates. During this epoch they diverged extensively with as many as 100 species and 14 genera roaming throughout the Old World (Begun, 2003). The familial relationships of the Miocene primates and the lineage which later led to the hominoid apes are at present highly debated as the (surprisingly abundant) fossil record displays highly variable morphology over an equally variable timescale. Prominent Miocene Apes include; the Ugandan *Morotopithecus* which is believed to be ancestral to modern day Gibbons (Young and MacLatchy, 2004) and whose appearance corresponds with the predicted molecular divergence of the Gibbons (Table 1.2); the Eurasian *Ramapithecus* or *Sivapithecus* which is ancestral to Orang-utans (Madar et al., 2002) which is present between 12.5 to 18 Mya, in accordance with molecular estimates of Orang-utan divergence (Table 1.2.); and *Gigantopithecus* which persisted in Asia until 0.5 Mya during the Pleistocene (Ciochon et al., 1996).

At the beginning of the Miocene the Old World Monkeys (Cercopithecoids) and Apes (Hominoids) diverged in East Africa with the primate *Proconsul* being a very early Ape and *Victoriapithecus* being an early ancestor of Old World monkeys

Table 1.2 Divergence Times of the Primates in Relation to the Human Lineage Based Upon Molecular Studies

Study	Method	NWM	OWM	Gibbon	Orang-utan	Gorilla	Chimp
(Sarich and Wilson, 1967)	Cross reactivity of Albumin	--	30*	--	--	--	5
(Sibley and Ahlquist, 1984)	DNA Hybridisation	--	27-33	18-22	13-16	8-10	6.3-7.7
(Sibley and Ahlquist, 1987)	DNA Hybridisation	--	25-34	16.4-23	12.2-17	7.7-11	5.5-7.7
(Hasegawa et al., 1987)	mtDNA and eta-globin	38*	25.3 ± 2.4	--	11.9 ± 1.7	5.9 ± 1.2	4.9 ± 1.2
(Arnason et al., 1996)	mtDNA (cyt b)	60	40	36	24.5	18	13.5
(Takahata and Satta, 1997)	Nuclear DNA	57.5	31	--	--	8	4.5
(Kumar and Hedges, 1998)	658 nuclear genes	47.6	23.3	14.6	8.2	6.7	5.5
(Stauffer et al., 2001)	9 nuclear genes	--	23*	14.9 ± 2.0	11.3 ± 1.3	6.4 ± 1.5	5.4 ± 1.1
(Chen and Li, 2001)	Range of coding and non-coding autosomal sequences	--	--	--	12-16*	6.2 ± 8.4	4.6 ± 1.2
(Glazko and Nei, 2003)	Nuclear genes and mtDNA	33	23	--	13 (12-15)	7(6-8)	5-7(6)
(Schrago and Russo, 2003)	mtDNA	35	25*	15-19	13-16	--	6-7
	Average	45.1	28.5	20.3	14.11	8.5	6.3

* Calibration point

(Benefit and McCrossin, 1997). During the Middle Miocene (17-10 Mya) the remnants of the Tethys Sea (the Mediterranean region today) sea dried up as a consequence of plate tectonics and climate change which facilitated the formation of a land bridge between Africa, Arabia and Eurasia. This permitted the range expansion of Apes into Eurasia who were presumably following food resources as the open woodlands of Africa were being replaced by savannah. The most prominent Eurasian Ape family from this time is *Dryopithecus* (13 to 8 Mya) which is found in many localities ranging from Spain to China (Moya-Sola and Kohler, 1996) and is currently believed to be a direct descendant of *Proconsul.*

During the period of 17 to 15 Mya there is currently very little evidence of the Apes in Africa, which signifies that the Apes evolved in Eurasia. However the Kenyan primate, *Kenyapithecus* (14 Mya), displays characteristics that indicate it is an ancestor of Chimpanzees and Gorillas, which demonstrates that the precursors of the human lineage could have been present on either continent (McCrossin and Benefit, 1993). It is most likely that the lineage which later gave rise to Gorillas, Chimpanzees and Humans evolved from the Dryopiths in Eurasia as the Greek hominoid *Ouranopithecus* (10 Mya), most closely resembles the hominid clade (de Bonis et al., 1990). This further implies that during the Late Miocene (10 to 5 Mya) the Apes returned to Africa, which is attributed to the spread of grasses and deciduous trees, an event termed the 'Vallesian Crisis' which occurred at approximately 9 Mya (Agusti et al., 2003). This change in vegetation occurred as a result of global cooling which was triggered by the expansion of the Antarctic ice sheet and consequently led to a rapid decline in the diversity and number of Apes

throughout the Old World as their habitats became restricted to regions along the equator. In contrast, this change in flora facilitated the radiation of the Monkeys.

Towards the end of the Miocene there is very little fossil evidence of the diversification and speciation of the Human, Chimpanzee and Gorilla lineages within Africa. Molecular evidence suggests that Gorillas split off between 6 to 11 Mya and Chimpanzees between 4.5 to 7.7 Mya (Table 1.2); these date estimations are variable as they are dependant upon the calibration point, the statistical method applied and the type of sequences utilised (Yoder and Yang, 2000). The divergence of Humans and Chimpanzees has recently been narrowed down to between 6 to 7 Mya as a result of the discovery of a partial skull (Toumai) with several jaw fragments and teeth in central Africa (Chad) (Brunet et al., 2002). These hominid fossils are representative of six individuals which have been given the species name *Sahelanthropus tchadensis.* According to the associated fauna, these hominid fossils are between 6 to 7 Mya which implies that the divergence between the human and chimpanzee lineages predates estimates by several molecular studies (Table 1.2) (Vignaud et al., 2002). A second late Miocene species, *Orrorin tugenensis*, found in Ethiopia, is estimated to be approximately 6 Mya and is claimed to be the first hominid of the human lineage (Pickford, 2001). However, it has been between suggested that 'Millenium Man' could be an early ancestor to the Chimpanzee lineage as a third late Miocene primate, *Ardipithecus kadabba* (5.2 to 5.8 Mya), also from Ethiopia, displays a more hominid-like morphology (Aiello and Collard, 2001; Pickford, 2001; Haile-Selassie et al., 2004).

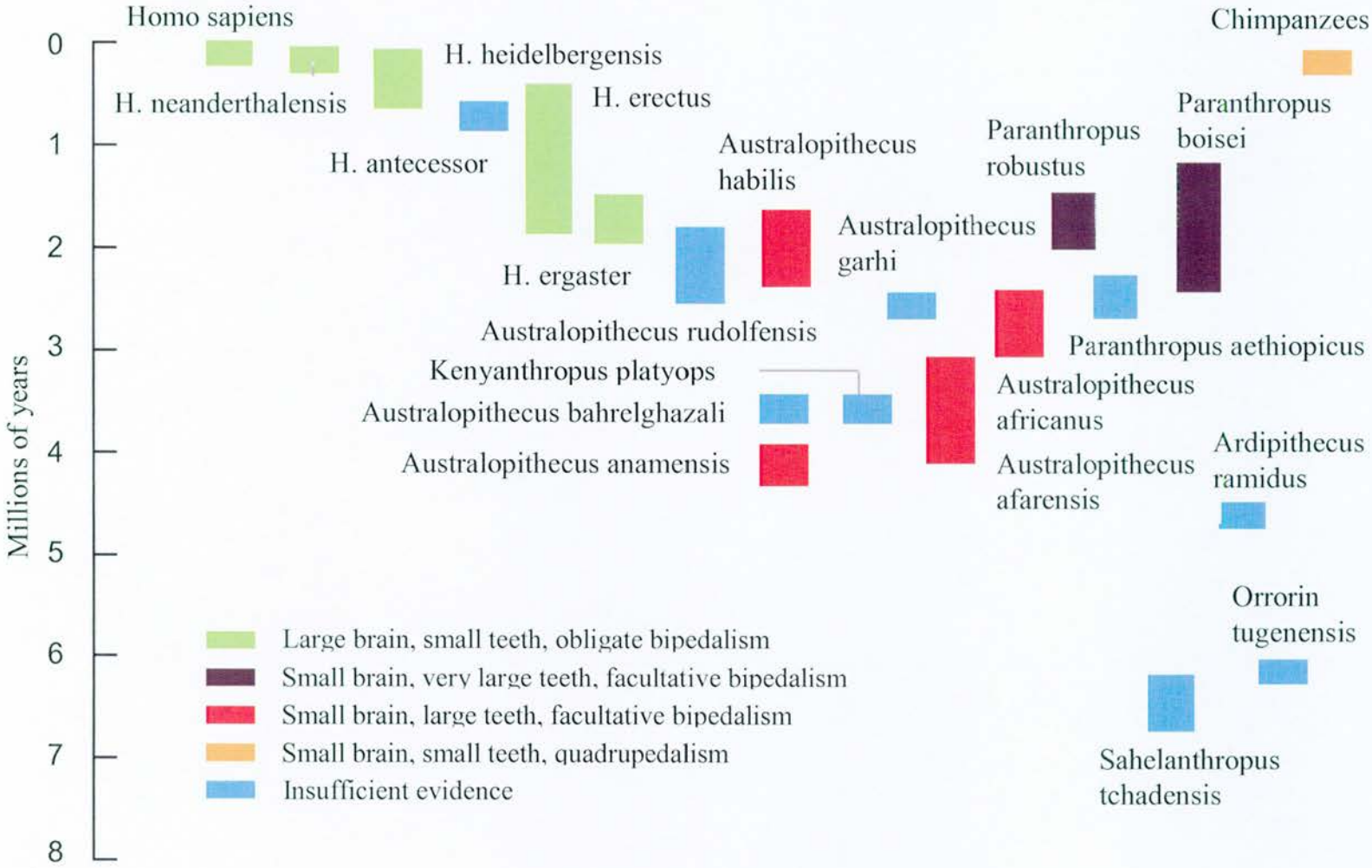
Primate evolution during the Pliocene epoch (1.8-5 Mya) is characterised by two major developments, the first, the diversification and spread of the monkeys and

the second, the evolution of the bipedal apes. Fluctuating climate during this epoch is presumed to have been an underlying cause of both of these developments; from the middle Pliocene (2.8 Mya) an Ice Age and subsequent glacial and inter-glacial periods ensued. The continents also finally reached their present-day positions and the Panamanian bridge formed joining North and South America between 2 to 3 Mya.

The earliest Pliocene bipedal primate, *Ardipithecus* (previously *Australopithecus*) *ramidus* was present in East Africa at 4.4 Mya and is presumed to be a member of the same species as *Ardipithecus kadabba* (White et al., 1994). The first true hominids are distinguished by their bipedalism and are generally classified into the genus *Australopithecus*. Throughout the Pliocene several species of australopithecines existed within Africa at the same time, with many exhibiting mosaic characteristics, this consequently thwarts an orderly evolutionary sequence between the forms (Figure 1.8).

The earliest known australopithecine, *Australopithecus anamensis*, lived in East Africa between 3.9 and 4.2 Mya and is believed to be the ancestor of all later bipedal hominids (Leakey et al., 1995). The next most prominent hominid within the mid Pliocene is *Australopithecus afarensis* of which the 'Lucy' skeleton is the most famous example. This species persisted in East Africa from 2.9 to 3.9 Mya and is believed to be a predecessor of both the 'Robust' and later 'Gracile' australopithecine forms which are distinguished on the basis of morphology related to diet. A second mid-Pliocene species has recently been discovered in East Africa which is estimated to have existed 3.5 Mya and is called *Kenyanthropus platyops*. It is contested as to whether it is an australopithecine as the cranium is badly distorted

Figure 1.8 Sequence of Hominids from the Pliocene to Holocene. Reproduced from (Wood, 2002)



(Leakey et al., 2001). Two gracile forms of australopithecine are proposed to be intermediate between *Australopithecus afarensis* and later Homo; the first is *Australopithecus gari* which was present in East Africa 2.5 Mya (Asfaw et al., 1999); and the second is *Australopithecus africanus* which existed in South Africa between 2.4 to 3 Mya. At this time robust australopithecines were also present in Africa, they are often referred to as the genus *Paranthropus* to distinguish them from the human lineage. *Paranthropus aethiopicus* represents the oldest form and was present in East Africa between 2.3 and 2.6 Mya. Aethopicus's descendants are *Paranthropus robustus* which persisted in South Africa between 1.5 to 2 Mya and *Paranthropus boisei* which lived in East Africa between 1.1 to 2.1 Mya (Alemseged et al., 2002).

The first hominid belonging to the genus Homo is the eastern African *Homo habilis* (handy man), which is observed across the Pliocene to Pleistocene boundary between 1.5 to 2.4 Mya. The first deliberate tool use is associated with *Homo habilis* with the tools referred to as the Oldowan tool culture. *Homo habilis* is a controversial species as it has been classified within both the genus Australopithecine and Homo; this is due to a high degree of morphological variation within the species. The definition of a second Homo species, *Homo rudolfensis*, has resolved some of classification issues (Miller, 2000; Blumenschine et al., 2003).

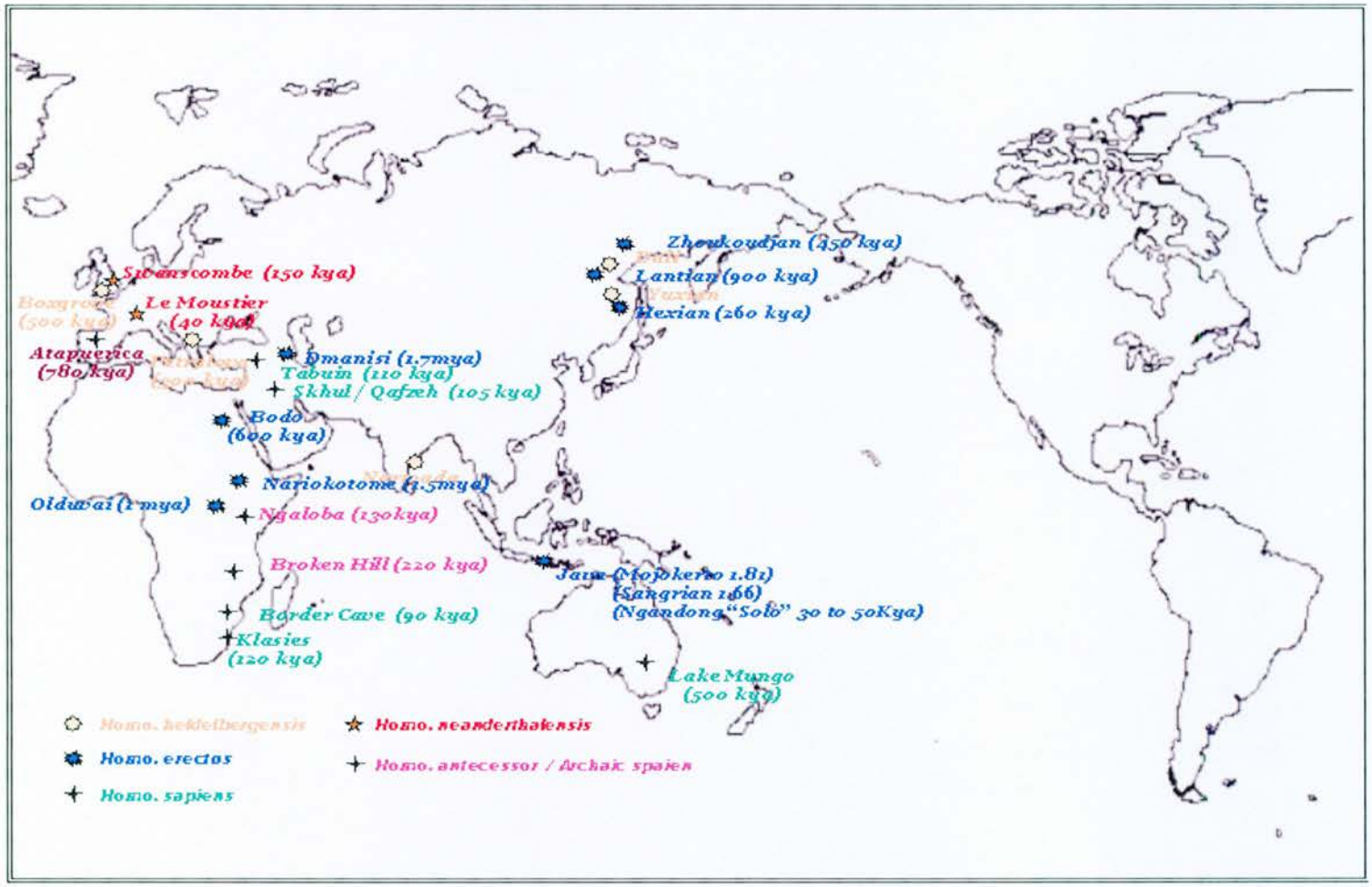
Until 2000, it was believed that the descendants of *Homo habilis*, referred to as *Homo erectus*, were the first hominids to leave Africa. Excavations in Dmanisi, Georgia disproved this presumption as several individuals who appeared to be intermediate between *Homo habilis* and *Homo erectus*, were discovered within strata which dated to 1.75 Mya (Vekua et al., 2002; Gabunia et al., 2000). These specimens appear to be more primitive than the African *Homo erectus*, which is frequently

referred to as *Homo ergaster* as the Eurasian *Homo erectus* is often regarded as a separate species from the African *Homo erectus.* As a result, the new Dmanisi species has been named *Homo georgicus*.

Following the exit of *Homo georgicus* an exodus of hominid forms ensued from Africa. Early Pleistocene examples are predominantly of the *Homo erectus* form which is often referred to as the 'globetrotter' as examples are found as wide-ranging as North Africa, Europe, and western Asia (Figure 1.9). Evidence of *Homo erectus* in Java between 1.6 and 1.8 Mya indicates that *Homo erectus* moved very quickly from Africa following its initial appearance at 1.9 Mya. As previously highlighted, it is debated as to whether the Asian and African forms of *Homo erectus* can be regarded as the same species (Kramer, 1993). However, it is accepted that the African form of *Homo erectus*, most notably exemplified by the 'Nariokotome Boy' at 1.5 Mya (Brown et al., 1985) and a more recent example found in Ethiopia existing at 1 Mya (Asfaw et al., 2002), are the ancestors of *Homo sapiens*. Late examples of *Homo erectus* within Eurasia are located at Ngandong, Indonesia at between 30,000 to 50,000 years ago (Swisher, III et al., 1996), although these dates are contested (Grun and Thorne, 1997).

The next distinct hominid species to be observed within the Pleistocene is *Homo antecessor* which has been found within the Spanish cave site of Atapuerca, dating to 780,000 years ago (Bermudez de Castro et al., 1997; Carbonell et al., 1999) and Ceprano, in Southern Italy where it persisted between 800,000 and 900,000 years ago (Manzi et al., 2001). Antecessor is the oldest European hominid and has been suggested to be the last common ancestor of *Homo sapiens* and *Homo neanderthalensis* (Carretero et al., 1999).

Figure 1.9 Hominid Sites Between 2 Million and 90,000 Years Ago



Homo heidelbergensis is often referred to as archaic *Homo sapiens* as it possess a morphology characteristic of both *Homo erectus* and modern humans. Examples are found between 200,000 and 500,000 years ago across Europe and the sub–continent (Roberts et al., 1994; Perez-Perez et al., 1999). The distinction between *Homo heidelbergensis* and *Homo erectus* is often confused as early *Homo heldelbergensis* exhibits a robust cranial morphology which is very similar to *Homo erectus* (Kramer, 1993)

Homo neanderthalensis existed between 30,000 and 230,000 years ago with examples located in the Middle East and Europe (Figure 1.9). Neanderthals latterly coexisted with modern humans (Cro-Magnons) in Europe and it is highly debated as to whether Neanderthals represent a sub-species of *Homo sapiens* and if the two hominid species interbred (Curnoe and Thorne, 2003). These issues are central to the models of modern human origins which will be discussed in Section 1.5.2.

1.5.2 Models of Human Origins

All models relating to the origin and evolutionary transition of *Homo sapiens* are based upon several critical questions, namely; when did *Homo sapiens* appear; where did they arise; and what is their relationship to the other hominid forms *Homo antecessor*, *Homo heidelbergensis* and *Homo neanderthalensis*? *Homo erectus* is accepted to be the ancestor of all later 'archaic' sapiens but the genetic relationships of all of the lineages derived from these later hominids are at present unresolved.

The initial model relating to the diversification of *Homo erectus* is referred to as the 'Multiregional Hypothesis'. This proposes that the evolutionary transformation of *Homo erectus* into *Homo sapiens* involved continuous change within a genetically coherent lineage, with the gene pool being maintained through significant inbreeding (Figure 1.10a.). Consequently, the genetic roots of modern humans are expected to be 'deep' as they should reflect the initial dispersal of *Homo erectus* from Africa (Wolpoff, 1996; Wolpoff et al., 2000).

A further assumption of the multiregional model is that the archaic *Homo sapiens* represent intermediate forms between *Homo erectus* and *Homo sapiens*. Central to the European facet of this model is the genetic relationship between *Homo sapiens* and *Homo neanderthalensis*. Assuming the model to be correct, the early forms of *Homo neanderthalensis* are expected to be ancestral to *Homo sapiens* and late Pleistocene members of both species interbred. Support for this model is provided by a child's skeleton in the Lapedo Valley of Portugal which is dated to 24,500 years BP. As the skeleton is present 4000 years after Neanderthals are last observed in Europe and it

Figure 1.10 Models of Human Origins during the Pleistocene

Multiregional

Europe Africa Asia

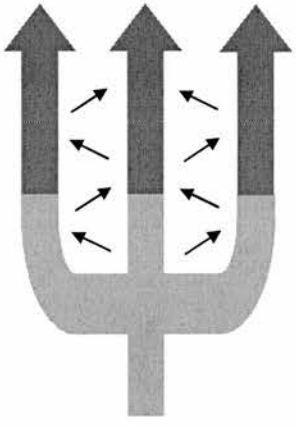


Figure 1.10a The Multiregional hypothesis proposes that throughout the Pleistocene genetic continuity has been maintained through significant gene flow and that the archaic forms of *Homo sapiens* are intermediate between *Homo erectus* and *Homo sapiens*.

Out of Africa

Europe Africa Asia

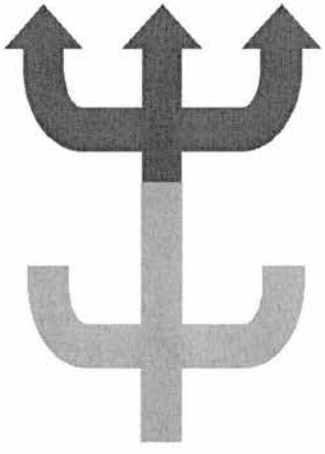


Figure 1.10b The Out of Africa hypothesis proposes that contemporary human populations have a recent, single origin, from Africa. The population expansion of *Homo sapiens* originating from Africa resulted in the replacement and subsequent extinction of archaic *Homo sapiens*.

possesses features of both Neanderthals and modern humans, it supports the view that *Homo neanderthalensis* was a subspecies of *Homo sapiens* (Duarte et al., 1999). A second individual who is represented by a jawbone dating to between 34,000 to 36,000 years BP, which was discovered within the Carpathian Mountains of Romania, also displays both Neanderthal and modern human traits and so reinforces the multiregional model (Trinkaus et al., 2003).

Genetic substantiation for a single evolving lineage of Pleistocene hominids has been the surveillance of the evolutionary history of a mitochondrial sequence which is present as a nuclear mitochondrial sequence (nmut) within contemporary humans. The extraction and amplification of the 'founder' mitochondrial sequence from the 60,000 years BP Australian 'Mungo Man' has been suggested to indicate that the mitochondrial DNA from this individual lies outside the range of modern humans and is strong evidence for the multiregional model (Adcock et al., 2001). This study has been highly criticised as the sequencing results were not independently duplicated (Cooper et al., 2001) and it has also been suggested that the acquired mitochondrial sequences had been subject to a high degree of post-mortem damage (Gilbert et al., 2003). In addition, the relative age of 'Mungo Man' has been contended (Bowler and Magee, 2000) and thermal history of the site makes it extremely unlikely that the DNA of the individual had survived (Smith et al., 2003).

Since the original proposal of the multiregional hypothesis, the hominid species *Homo antecessor* and *Homo heidelbergensis* have been identified throughout Africa and Europe. This has led to the adaptation of the multiregional model which is presently referred to as the 'Regional Continuity Model'; this proposes continuous evolution over the past one million years with racial variation developing early and the gene pool being

maintained through with significant gene flow (Walter et al., 2000). The genetic status of *Homo heidelbergensis* in relation to later hominids is unresolved, several hypotheses have been put forward. The first denotes that *Homo heidelbergensis* is ancestral to both *Homo neanderthalensis* and *Homo sapiens*. The second proposes that two species of *Homo heidelbergensis* existed at the same time, one within Africa and one within Europe, with the African species giving rise to *Homo sapiens*. The third hypothesis lends support to the regional continuity model where *Homo antecessor* and *Homo heidelbergensis* are part of a single evolving lineage. A forth suggestion is that *Homo antecessor* is ancestral to *Homo sapiens* and *Homo heidelbergensis* is ancestral to *Homo neanderthalensis*.

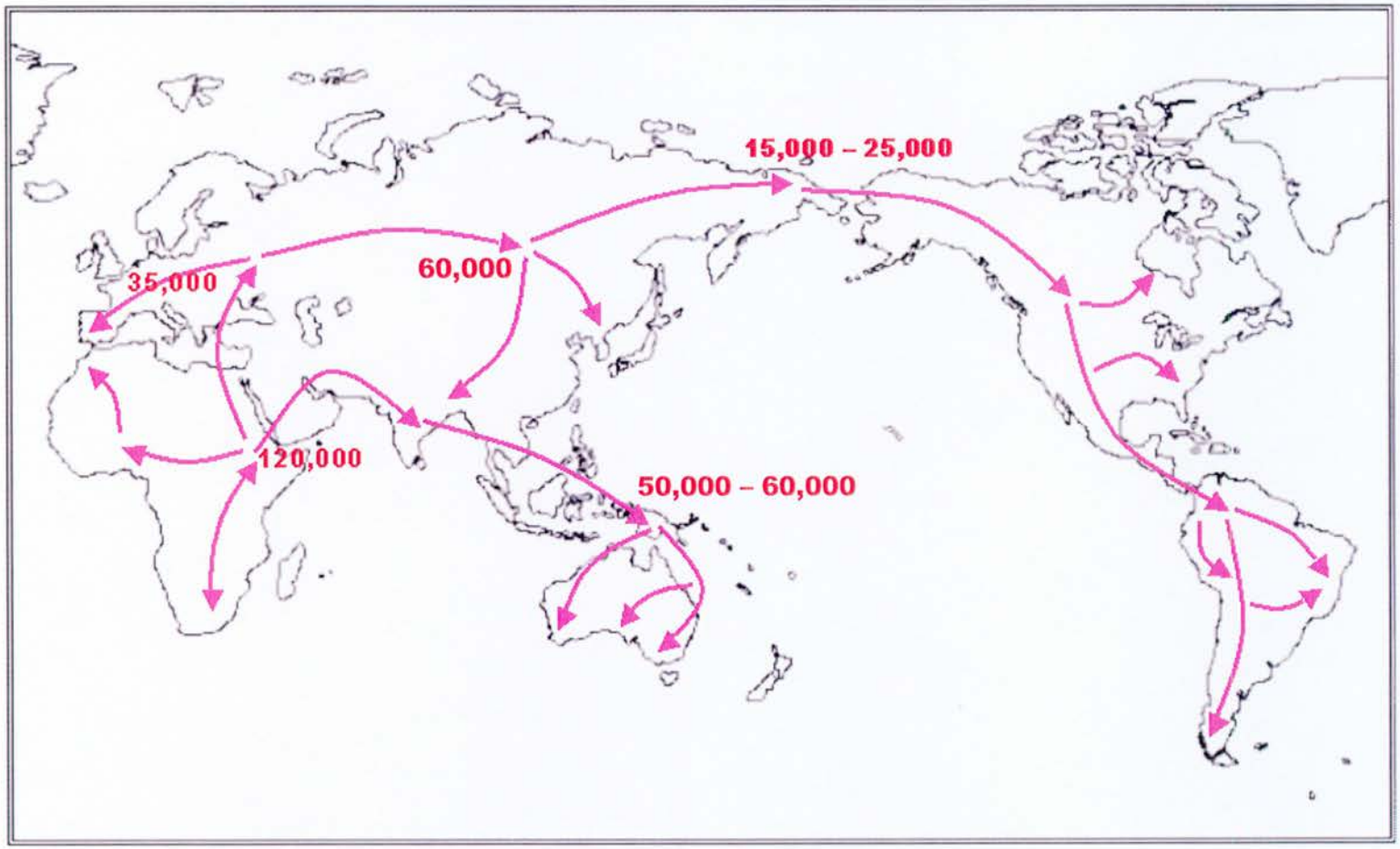
The successful amplification of mitochondrial DNA (mtDNA) from eight representatives of *Homo neanderthalensis*, who were dispersed both in time period and geography, has resolved some of the issues surrounding the origin of *Homo sapiens* (Krings et al., 1997; Ovchinnikov et al., 2000; Krings et al., 1999; Krings et al., 2000; Schmitz et al., 2002; Serre et al., 2004). In all eight cases the mtDNA appears to be distinct from any modern human lineage, indicating that extensive admixture between the two hominid species did not occur (Serre et al., 2004). Included within this data set, are two Neanderthal individuals from Vindija, Croatia who, based upon morphology, have been described as 'intermediate' or 'transitional' Neanderthal forms (Smith et al., 1985). In addition, comparison of mtDNA belonging to seven Cro-Magnons to contemporary human mtDNA has indicated that the Cro-Magnons fall within the expected range of present day human populations (Caramelli et al., 2003; Serre et al., 2004). These results suggest that an alternative model to the multiregional model,

termed the 'Out of Africa' or 'Regional Replacement', is a more viable model for modern human origins.

The 'Out of Africa' model proposes that contemporary human populations have a single recent origin within Africa, with initial appearance approximately 200,000 years ago during the Middle Paleolithic (Figure 1.10b). In this model, regional morphological differences developed within the last 100,000 years, in succession from the dispersal of anatomically modern humans from Africa (Figure 1.11), The dispersal of contemporary *Homo sapiens* is posited to have began approximately 100,000 years ago, with appearance in the eastern Mediterranean approximately 90,000 years ago with occupation of Europe and Asia by at least 40,000 years ago. The region of Sahul, which was purportedly reached due to the lowering of sea levels and is represented by the islands Papua-New Guinea and Australia today, is believed to have been colonised between 50,000 to 60,000 years ago (Redd and Stoneking, 1999). It is uncertain when exactly humans first left North-eastern Asia to cross the Bering Strait, this was presumably achieved by the passage across a land bridge generated from a combination of ice and dispersed islands, to enter North America. The earliest secure dates for the occupation of America are around 14,000 years ago, recent evidence suggests that colonisation could have been much earlier due to the presence of human in the Arctic circle at 30,000 years ago (Pitulko et al., 2004). With the exception of Polynesia which was first colonised 3000 years ago, by 10,000 years ago most of the land areas of the world were subsequently occupied.

Definitive forms of *Homo sapiens* have been observed at the Klasies river in South Africa (Deacon, 1992), which are estimated to be between 125,000 to 130,000 years BP, and also at Omo Kibish in Ethiopia (Bartsiokas, 2002), which are less than

Figure 1.11 Global Movement of Anatomically Modern Humans in the last 120,000 years



130,000 years old. Recently, three skulls dating to 160,000 years BP, were discovered in Herto, Ethiopia. They have been assigned to the subspecies *Homo sapiens idaltu* as they represent an intermediate form of *Homo sapiens* with anatomy that is not fully modern (White et al., 2003). The anatomy of the Herto skulls is regarded as definitive evidence that modern humans originated in Africa, although the designation of a new subspecies is contended (Stringer, 2003).

A further conjecture of the 'Out of Africa' model is that the population expansion of *Homo sapiens* resulted in the replacement and subsequent extinction of *Homo neanderthalensis* in Europe. Extensive genetic admixture between the two species of hominids in Europe is therefore not accepted to have occurred. Reinforcement for this view has been provided by the seminal study of blood groups (Cavalli-Sforza et al., 1994) and a wide range of genetic studies including; mtDNA (Cann et al., 1987), X chromosome (Zietkiewicz et al., 1997; Harris and Hey, 1999; Kaessmann et al., 1999), nuclear genes (Harding et al., 1997; Nickerson et al., 1998), non-coding autosomal regions (Zhao et al., 2000; Yu et al., 2001), and the Y chromosome (Huang et al., 1998; Semino et al., 2000). With the exception of the autosomal regions, all these studies show that contemporary human populations are not panmictic. While the Mitochondrial sequence data retrieved from eight Neanderthals and seven Cro-Magnons (described above) also lends support to the out of Africa model. The analysis of MtDNA only reflects the history of the maternal lineage and recent reports of heteroplasmy and recombination within mitochondrial genomes are of concern when constructing a phylogeny (Awadalla et al., 1999; Eyre-Walker and Awadalla, 2001). Statistical validation of the genetic admixture or isolation of Neanderthals and Cro-Magnons during the Pleistocene is either dependent upon the

analysis of many more Pleistocene hominids (Wall, 2000) or the analysis of at least 50 unlinked nuclear loci within more than 10 contemporary Europeans (Nordborg, 1998).

An extension to the 'Out of Africa' model is the 'Assimilation Model'. This advocates that *Homo sapiens* primarily arose in Africa but that they were subject to a complex and long-term process of interbreeding and population movement (Smith, 1985; Yu et al., 2001). Recent re-analysis of the demography of human DNA sequence data using Nested Clade Analysis (Cann, 2002; Templeton, 2002) and population based analysis of coding and non-coding autosomal regions (Harding et al., 1997; Zhao et al., 2000; Yu et al., 2001), indicate that the 'Assimilation Model' offers a more reliable account of human dispersal than the 'Out of Africa' or 'Multiregional' models.

CHAPTER 2

MATERIALS AND METHODS

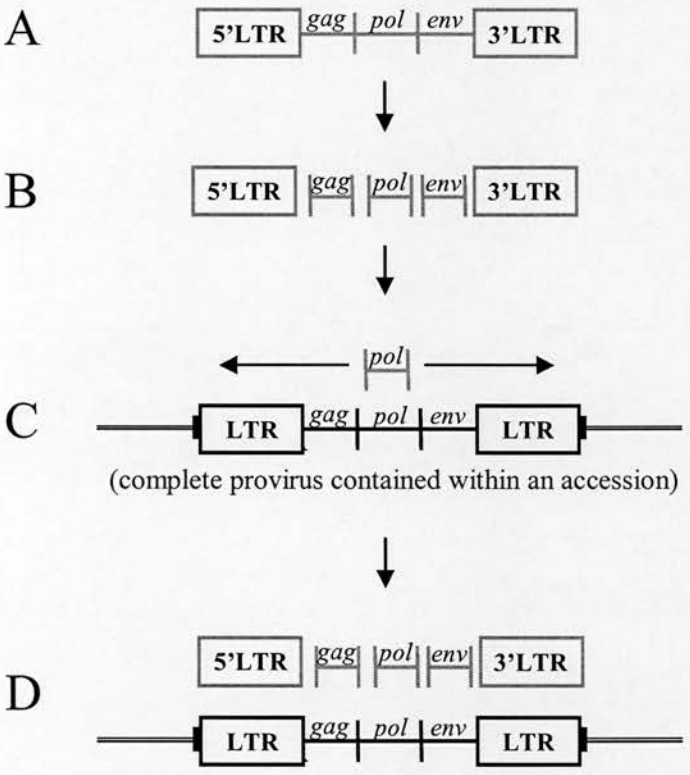
2.1 Data Mining for HERV-K and other Retrotransposons

2.1.1 Basic Local Alignment Search Tool (BLAST)

To determine HERV-K sequences and related retrotransposons within the human genome, *in silico* searches were conducted of the redundant and non-redundant nucleotide sequence databases utilising the Basic Local Alignment Search Tool (BLAST). Details of the sequence databases are provided in Section 2.1.2. BLAST is a heuristic programme, which performs local alignments as opposed to global alignments. This offers the advantage of being able to find more similarities in homologous sequence accessions which contain large scale indels that interrupt similarity to the query sequence.

When scoring sequences within a database, the BLAST programme builds a position specific matrix for high scoring accessions (Altschul et al., 1997) (Figure 2.1). This is achieved by breaking the query sequence into short fragments and then searching the (nucleotide) sequence databases for identical or close matches. Following the determination of an equivalent sequence (hit), a local alignment is generated by extending the closely matching sequence in both directions. The quality of each alignment is then statistically examined to provide an indication of similarity to the query sequence. This is achieved by calculating the probability of the alignment arising by chance, which is referred to as the expected value (E - value). An alignment with a low E – value (close to zero) is less likely to have occurred by chance and has a greater statistical significance than an alignment that has a

Figure 2.1 Description of the BLAST for HERV Proviruses



Query sequence

Query proviral sequence was broken up into fragments to be used as probes to find short homologous sequences or ' hits'.

Each hit was then expanded to generate a paired alignment.

The optimum alignment for each query sequence pair was then assembled and each pair was assigned an E - value, giving an indication of the statistical significance of the alignment. High scoring accessions containing homologous HERV sequences, had a lower E - value.

higher E – value. For example, a sequence alignment with an E - value of 0.05 indicates that this similarity has a probability of 5 in 100 (1 in 20) of occurring by chance alone. Alignment pairs that have low E – values are referred to as high scoring segment pairs (HSPs). Sequence accessions that contain HSPs are presented at the top of the BLAST result output.

A further advantage of the BLAST is that within a results output the position a HSPs within a sequence accession are listed which facilitates further sequence analysis. However, a disadvantage is that only regions of high homology to the query sequence are listed, the internal or flanking regions are not retained. To analyse these further, high scoring accessions must be examined individually.

2.1.2 HERV-K Proviral Genomes

To identify and locate HERV-K proviral sequences belonging to the HML-2, HML-3 and HML-4 subgroups; the GenBank non-redundant and high-throughput genomic sequence database (http://www.ncbi.nlm.nih.gov/genome/seq/HsBlast.html), the Ensembl (human) database (http://www.ensembl.org) and the HERV-d (http://herv.img.cas.cz) database were screened using the BLAST. The sequence accessions M14123 and AF020092 were used as queries for the HERV-K(HML-2) and HERV-K(HML-4) subgroups as they were the published consensus sequences. For the HERV-K(HML-3) subgroup a consensus proviral sequence was generated by hand in the SIMMONIC sequence analysis package (http://www.polio.vir.gla.ac.uk/)

(Simmonds et al., 2004) as a representative sequence submission did not exist within any of the sequence databases.

Throughout the duration of this study, the nucleotide sequence databases were periodically examined in order to update information from subsequent releases. Sequentially, the NCBI releases of the human genome that were screened were; 24 (24th December 2001), 28 (24th June 2002), 31 (November 2002) and 33 (14th April 2003).

Following the completion of relevant BLAST searches, each individual high scoring accession was imported and aligned by hand within the SIMMONIC sequence analysis package. This allowed the retention of 1500 bp of cellular sequence that flanked each of the putative proviruses. Individual proviruses were then identified on the basis of their cellular flanking sequences and chromosomal location. Where an indel in the form of an insertion was present within a proviral sequence, the insertion was removed and entered as a BLAST query into the nucleotide sequence databases in order to determine its origin.

Alignment of sequence entries belonging to the three HERV-K subfamilies was carried out following each of the published consensus sequences. As many of the sequence entries appeared to be highly divergent, where putative ORFs were present, the alignment was followed at the amino acid level in order to maintain a sensible alignment.

2.1.3 Construction of SVA and LINE Retrotransposon Datasets

To construct datasets of LINE and SVA retrotransposons, the GenBank non-redundant and high-throughput genomic sequence databases were screened by means of the BLAST programme using the accessions; U93573 (contained within AL354951) to search for LINE elements (Sassaman et al., 1997) and the SVA_{STPA1} retrotransposon (contained within AC016142) to search for SVA elements (Ostertag et al., 2003). To obtain a large number of sequences, an option of 500 alignments for LINE elements and 100 alignments for SVA elements was requested. The BLAST results produced 10760 'hits' for the LINE query and 15073 'hits' for the SVA query.

Following this, the BLAST results were saved as .cgi files and imported into the SIMMONIC sequence analysis package. To exclude the 'hits' that did not constitute equivalent complete query sequence, all sequences that were not full length were removed from the alignment. This left the separate totals of 527 LINE sequences and 92 SVA sequences. Where possible, sequences were then aligned at the amino acid level in order to maintain open reading frames (ORFs). Identical sequences were then removed from the alignment to ensure that either a LINE or SVA retrotransposon was not represented more than once. This left the totals of 327 LINE elements and 89 SVA elements. Following this, all sequences that shared 1 % nucleotide similarity were also excluded, ensuring that the datasets consisted of divergent retrotransposons. This left the total of 192 LINE elements and 44 SVA elements within the datasets.

2.2 Collection and Extraction of Genomic DNA

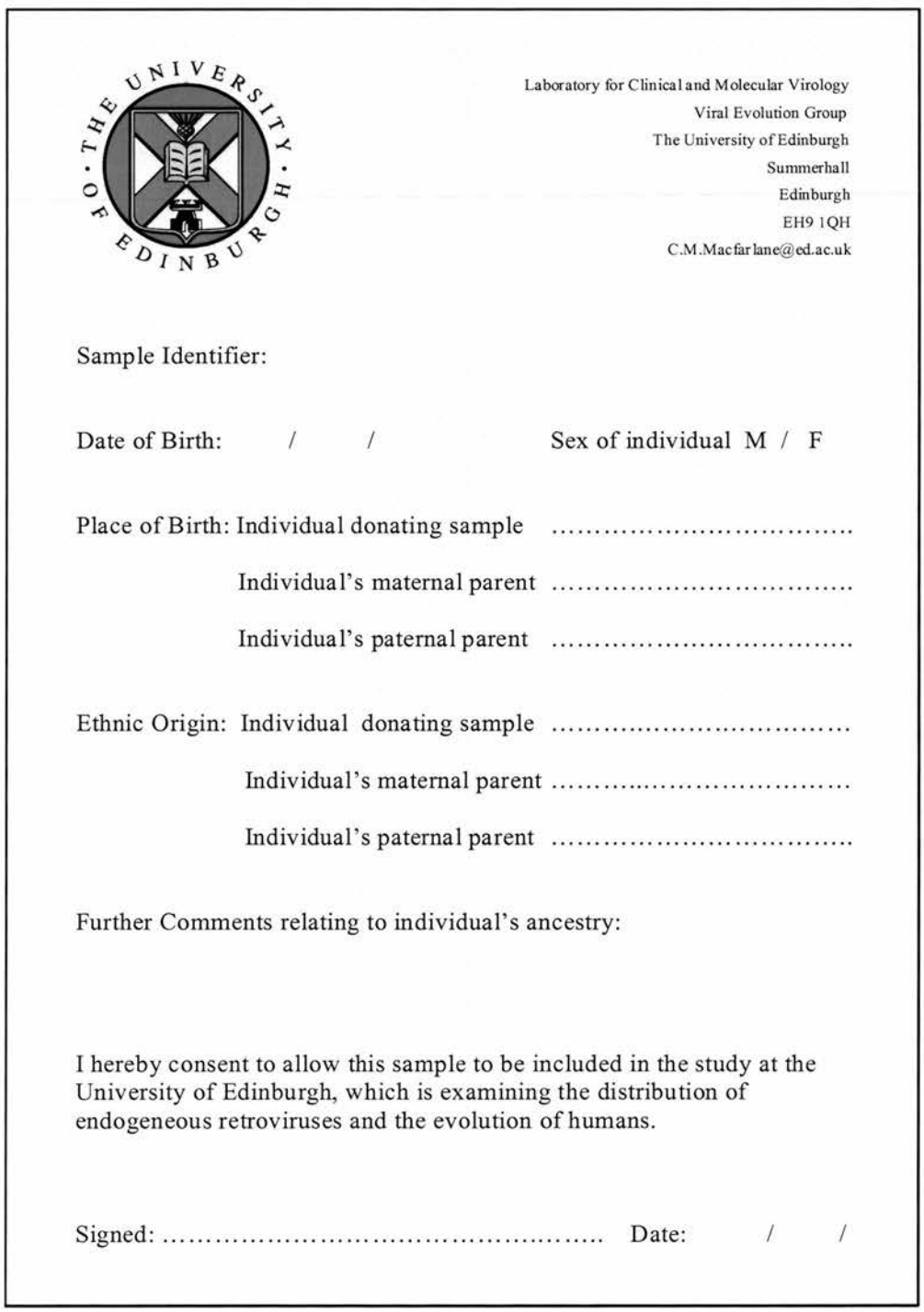
2.2.1 Collection of Genomic DNA Samples

Serum samples were provided by the Scottish National Blood Transfusion Service (SNBTS) and were originally collected for the epidemiological investigation of Hepatitis B Virus (HBV) transmission. Each sample collected was from a 'Highland' Papua-New-Guinean. Primate buccal swabs were kindly taken by Anna Meredith and came from primates located at Edinburgh Zoo. The remaining human DNA samples were collected specifically for this study and were obtained in the form of buccal swabs.

To assist in the critical evaluation of single populations, individuals were asked to provide information regarding their population affiliation and gender (Figure 2.1) (Appendix B, Table B.1). However, as each sample was allocated a unique number identifier, the samples remained anonymous.

As many of the buccal swabs were to be collected from a hot, tropical environment, experimental samples belonging to the same individual (male) were stored at 37 ^{o}C and left for a three week period to analyse DNA survivability. Half the samples was stored as 'dry' in their collection tubes and the remaining were stored as 'wet' in 1 ml of 80 % (w/v) Ethylene-diamine-tetra-acetate (EDTA) (pH 8). Nucleic acid integrity of single and high copy number DNA was assessed by polymerase chain reaction (PCR) amplification for the amelogenin gene and mitochondrial HV1 region (Section 2.3.2). The results indicated that storage in 1 ml

Figure 2.2 Collection of Anthropological Variables. In conjunction with the donation of a buccal swab, each sample donor was asked fill in the following form. As each sample was allocated a unique number identifier, the samples remained anonymous.



Laboratory for Clinical and Molecular Virology
Viral Evolution Group
The University of Edinburgh
Summerhall
Edinburgh
EH9 1QH
C.M.Macfarlane@ed.ac.uk

Sample Identifier:

Date of Birth: / / Sex of individual M / F

Place of Birth: Individual donating sample .....

Individual's maternal parent .....

Individual's paternal parent .....

Ethnic Origin: Individual donating sample .....

Individual's maternal parent .....

Individual's paternal parent .....

Further Comments relating to individual's ancestry:

I hereby consent to allow this sample to be included in the study at the University of Edinburgh, which is examining the distribution of endogeneous retroviruses and the evolution of humans.

Signed: ..... Date: / /

of 80 % EDTA (w/v) (pH 8) provided better long term storage for transport than dry preservation (Section 4.2.2) and was thus used for all buccal swab collections.

To ensure the integrity of a buccal swab, individuals were requested to have not eaten for at least 30 minutes prior to removal of cheek cells. Removal was achieved by vigorously scraping the cheek at least six times with a sterile (Wattman) cotton swab. The swab was then replaced in its sterile collection tube and 1 ml of 80 % EDTA (w/v) (pH 8) was added. EDTA was provided in a sterile 1.5 ml screw top eppendorf to prevent contamination of the sample. Each buccal swab was then sealed with parafilm tape to prevent leakage. In cases where an individual was not collecting a buccal swab of themselves; sterile gloves were worn throughout the procedure to prevent cross contamination of the sample.

2.2.2 Extraction of Genomic DNA from Buccal Swabs and Serum

Genomic DNA was extracted and purified from both buccal swab and serum samples using the QIAamp DNA extraction kit (QIAGEN, UK) following the manufacturers instructions and buffers provided (unless otherwise stated). In this protocol a mixture of detergent, guanidinium salt and protease is used to lyse collected cells thus releasing genomic DNA. QIAamp extraction columns contain a silica-gel membrane which adsorbs DNA in high salt conditions and releases it in lower salt concentrations. The high salt concentration of the lysate buffer is optimal for DNA adsorption to the silica-gel membrane within QIAamp spin columns. Absorption of DNA to the silica is also facilitated by the presence of guanidinium

salt which acts as a chaotropic agent. Bound DNA is then washed several times to remove cellular protein debris and inhibitors. Excess salts are then removed, allowing elution of bound DNA in 10 mM Tris-Cl pH 8.5. As the lysis procedures for the buccal swabs and serum samples varied, they will be considered separately below.

Prior to cell lysis, the tip of the cotton buccal swab was separated from the stick with scissors and placed in a microcentrifuge tube. 200 μl of the 80 % EDTA (v/v) (Ph 8) preservative from the buccal swab collection tube and 200 μl of Phosphate Buffered Saline (PBS) (Sigma) was then added to the microcentrifuge tube. Cells were then lysed by the addition of 20 μl of QIAGEN protease and 400 μl of buffer AL and mixed immediately and thoroughly. Buffer AL contains detergent, which along with the protease results in cell lysis and also increases the salt concentration of the lysate. Samples were then incubated at 56 ^{o}C for 10 min and briefly centrifuged to remove drops from the inside of the lid. 400 μl of 100 % ethanol (v/v) was then added and the samples vortexed and again briefly centrifuged. 700 μl of the lysate was then added to QIAamp spin column and genomic DNA was bound to the silica-gel membrane in the column by centrifugation at 8000 rpm for 1 min. Each spin column was then carefully transferred to a fresh collection tube and the flow through discarded. This process was then repeated with the final 520 μl of lysate. Bound genomic DNA was then washed by the addition of 500 μl of buffer AW1 with centrifugation at 15000 rpm for 1 min. This acted to remove endonucleases which may have been released during cell lysis. As performed before, each the spin column was placed in a clean microcentrifuge tube and the flow through discarded. Bound genomic DNA was further washed by the addition of 500

μl of buffer AW2 with centrifugation at 8000 rpm for 1 min. After transferring each spin column and discarding the flow through, the spin columns were centrifuged at 8000 rpm for 1 min to remove any residual buffer. Each spin column was then placed in a clean 1.5 ml collection tube and 200 μl of elution buffer was carefully added to the centre of the silica-membrane. The lids of the spin columns were shut and the samples left on the bench for 10 mins at room temperature to facilitate release of bound DNA. Genomic DNA was then eluted by centrifugation at 15000 rpm for 1 min. This process was then repeated using the 200 μl of elute to increase the yield of DNA. Following this second elution step, DNA samples were aliquoted in to volumes of 50 μl and stored separately. Per sample, this reduced the amount of DNA that was exposed to freezing / thawing during the conduction of PCR and so assisted the preservation of high quality DNA samples.

The lysis steps for extraction of genomic DNA from serum samples differed in a number of ways from the above protocol. Sample volumes were made up to 200 μl with PBS and then added to a microcentrifuge tube containing 20 μl of QIAGEN protease. After the addition of 200 μl of buffer AL, the samples were mixed thoroughly and incubated at 56 ^{o}C for 10 min. 200 μl of 100 % ethanol (v/v) was then added and samples vortexed. Lysate was then added to the QIAamp columns and the genomic DNA adsorbed to the silica gel by centrifugation at 8000 rpm for 1 min. The same protocol for washing and eluting bound DNA from buccal swabs was followed.

2.3 Polymerase Chain Reaction and Automated Sequencing

2.3.1 Polymerase Chain Reaction

Polymerase chain reaction (PCR) is a technique for the *in vitro* amplification of specific DNA sequences by the simultaneous primer extension of complementary strands of DNA. The sensitivity of PCR can be reduced when using low copy number or degraded DNA template, in which case nested or hemi-nested PCR can be used. These techniques utilise two consecutive rounds of PCR amplification. For nested amplifications, the first round (primary reaction) contains an external pair of primers while the second round (secondary reaction) contains two primers that are internal to the first primer pair. Hemi-nested reactions vary from nested reactions during the second round of amplification. Here, one of the primers used in the first round of amplification and a single nested primer which is internal to the first primer pair are used. In both methods, the amplicon produced by the first round of PCR is used as a template for the second round PCR amplification. The use of a second round of PCR amplification also increases the specificity of the technique, allowing the differentiation of a large number of similar, yet polymorphic sequences.

2.3.2 Primers and Conditions

Single round and primary PCR reactions were carried out in volumes of 50 μl with each containing; 200 - 400 ng of extracted genomic DNA, 200 μM each of dGTP, dATP, dTTP and dCTP, 0.5 μM of each outer primer (sense and antisense primers, respectively) and 0.5 Units of *Taq* polymerase in standard PCR buffer (Promega). The PCR reaction buffer (Promega) contained; 50 mM KCl, 10 mM Tris-HCl (pH 9.0), 0.1% Triton X-100 and 1.5 mM $MgCl_2$. Each sample was covered with a drop of liquid paraffin oil and transferred to a Techne Genius thermal cycler.

Secondary PCR reactions were performed in volumes of 30 μl, using 2 μl of the primary PCR reaction as the DNA template. 0.2 μM of each inner primer was used with the reaction mix as listed above.

Throughout the course of this study, known positive and negative controls were included within each set of PCR amplifications to check the integrity of the reaction and ensure that there was no DNA contamination.

A complete listing of; regions amplified, primer sequences, template denaturation, primer annealing and strand elongation conditions, for each PCR amplification, are shown below in Tables 2.1 to 2.4. With the exception of the 'Amelogenin' PCR amplification (Faerman et al., 1995), all primers and conditions are unique to this study.

Table 2.1 Primers and Conditions used for PCR Amplification to Confirm the Authenticity of a Sample. (1°) refers to primers used in primary PCR amplification and (2°) refers to nested primers used in the secondary PCR amplification. [a] Described in Faerman et al., (1995).

Region Amplified	Name	Orientation	Sequence (5'-3')	Cycle (seconds)	Size (bp)
Human specific Amelogenin X or Y chromosome [a]	AMG M4 AMG M5 AMG M6	S AS S	CAGCTTCCCAGTTTAAGCTT TCTCCTATACCACTTAGTCACT GCCCAAAGTTAGTAATTTTACCT	35 [94-60; 50-60; 72-60]	330 218
Human specific Mitochondria HVS2	PMT1 PMT2	S AS	CTCACCCATCAACAACCGCTAT GGGAGCAGAAGGGATTTGACTG	30 [94-60; 55-60; 72-60]	299
Mammalian Mitochondria Cytb	CytbS CytbAS	S AS	CACCCYTAYTACACMATYAAAGA GGGRTRTARTTGTCTGGGTCKCC	35 [94-60; 55-60; 72-60]	112
Primate Mitochondria HVS1	APEHV1S APEHV1AS	S AS	CATCWCGATGGATCACRGG TAAAARTGCATACCGCC	30 [94-60; 55-60; 72-60]	400
Primate Mitochondria 12S rRNA	Mt641 Mt1298	S AS	CCATAAACAMAYAGGYTTGGTCC CAGGGTTTGCTGAAGATGGCGGTATATA	35 [94-60; 55-60; 72-60]	657
Primate Protamine Gene (Nested PCR)	(1°)PRMS (1°)PRMAS (2°)PRM3 (2°)PRM4	S AS S AS	AGGTACAGATGCTGYCGCAG TCAGGCAGGAGTTTGGTGGA AGCCAGAGCCRGAGCAGAT CCTCAGCTGGGCCCACTTA	30 [94-60; 58-60; 72-60] 30 [94-60; 58-60; 72-60]	284 108

Table 2.2 Primers and Conditions used for PCR Amplification to Confirm The Relative Age of HERV-K(HML-2) Proviruses.

Region Amplified	Name	Orientation	Sequence (5'-3')	Cycle (seconds)	Size (bp)
K101 Proviral LTR	K101 PROKLTR	S AS	GAGTTATTAAGCGCAATCTTCTG AATGGAGTCTGGYATGTCTACT	30 [94-60; 50-60; 72-60]	375
K102 Proviral LTR	K102 PROKLTR	S AS	TCTCCACTTGAAGTGGTCATACT AATGGAGTCTGGYATGTCTACT	30 [94-60; 58-60; 72-60]	579
K103 Proviral LTR	5K103 PROKLTR	S AS	CCACCATCTGAGAAGTGTGATG AATGGAGTCTCCYATGTCTACT	35 [94-60; 60-60; 72-60] 72-360	235
K104 Proviral LTR	K104 PROKLTR2	S AS	CATATAAACAGAACTGTGGGGAA CAGTCTATAGATGTGGATGCCT	94-120; 35 [94-30; 58-30; 72-30] 72-360	278
K106 Proviral LTR	5K106 PROKLTR	S AS	TCCACCTGCGGACCTCCTCT AATGGAGTCTCCYATGTCTACT	94-120; 35 [94-30; 58-30; 72-30] 72-360	220
K107 Proviral LTR	5K107 PROKLTR	S AS	GGACACCCAACCTGCATGGT AATGGAGTCTGGYATGTCTACT	94-120; 35 [94-30; 55-30; 72-30] 72-360	192
K108 Proviral LTR	5K108 PROKLTR2	S AS	GTTACAGGAGTGCGCCATCAC CAGTCTATAGATGTGGATGCCT	94-120; 35 [94-30; 58-30; 72-30] 72-360	239
K109 Proviral LTR	K109 PROKLTR	S AS	CATCATGCTTAGAATACACCTATC AATGGAGTCTGGYATGTCTACT	30 [94-60; 58-60; 72-60]	430
12q14.1 Proviral LTR	AC025420S PROKLTR2	S AS	ACGTGCTGACCACTGGTGAG CAGTCTATAGATGTGGATGCCT	94-120; 35 [94-30; 58-30; 72-30] 72-360	345
11q22.1 Proviral LTR	AP000776 PROKLTR	S AS	TCATGTCTAGTGTATCTGATTCTC AATGGAGTCTGGYATGTCTACT	30 [94-60; 58-60; 72-60]	295
3q27.2 Proviral LTR	AC069420AS PROKLTR2	S AS	TGAGACAGGTACATGTGGGGAA CAGTCTATAGATGTGGATGCCT	94-120; 35 [94-30; 58-30; 72-30] 72-360	278

Table 2.2 ***Continued.*** Primers and Conditions used for PCR Amplification to Confirm The Relative Age of HERV-K(HML-2) Proviruses. (1°) refers to primers used in primary PCR amplification and (2°) refers to nested primers used in the secondary PCR amplification.

Region Amplified	Name	Orientation	Sequence (5'-3')	Cycle (seconds)	Size (bp)
11q23.2 Proviral LTR (Hemi-nested PCR)	(1°)AP000831	S	TCTGCTCCCAATGCAACTCAT	94-120; 35 [94-30;	
	PROKLTR2	AS	AGGGMGTRGTGATGACTCTTAA	55-30; 72-30] 72-360	446
	(2°)AP000831	S	TCTGCTCCCAATGCAACTCAT	94-120; 35 [94-30;	
	PROKLTR	AS	AATGGAGTCTGGYATGTCTACT	55-30; 72-30] 72-360	255
10p14 Proviral LTR (Hemi-nested PCR)	(1°)AC015686	S	TGCTGATGCATTACCTGCAGA	94-120; 35 [94-30;	
	PROKLTR2	AS	AGGGMGTRGTGATGACTCTTAA	55-30; 72-30] 72-360	324
	(2°)AC015686	S	TGCTGATGCATTACCTGCAGA	94-120; 35 [94-30;	
	PROKLTR	AS	AATGGAGTCTGGYATGTCTACT	55-30; 72-30] 72-360	133
3p25 Proviral LTR	53p25	A	CCGAGCTCTGTTGCACATGA	94-120; 35 [94-30;	240
	PROKLTR	AS	AATGGAGTCTGGYATGTCTACT	58-30; 72-30] 72-360	
19p13.11a Proviral LTR	519p13	A	GTACTCATACACACTGACCAG	94-120; 35[94-30;	289
	PROKLTR	AS	AATGGAGTCTGGYATGTCTACT	58-30; 72-30] 72-360	

Table 2.3 Primers and Conditions used for PCR Amplification to Confirm HERV-K(HML-2) Allelic Variants.

Region Amplified	Name	Orientation	Sequence (5'-3')	Cycle (seconds)	Size (bp)
K113 Insertion site	5K113 3K113	S AS	TGCATGGGGAGATTCAGAACC ATCCATACATTTCTGAGTCCTGA	94-120; 35 [94-30; 56-30; 72-30] 72-360	371
K113 Proviral LTR	5K113 PROKLTR	S AS	TGCATGGGGAGATTCAGAACC AATGGAGTCTCCYATGTCTACT	94-120; 35 [94-30; 56-30; 72-30] 72-360	273
K113 Full Provirus	5K113 GAG	S AS	TGCATGGGGAGATTCAGAACC GGATCTCTYGTCGACTTGTCC	95-120; 35 [95-30; 58-30; 72-90] 72-360	1258
K115 Insertion site	5K115 3K115	S AS	AGCACTGAGATCCAAACTCATAT CAGTCTATAGATGTGGATGCCT	94-120; 35 [94-30; 58-30; 72-30] 72-360	223
K115 Proviral LTR	5K115 PROKLTR2	S AS	AGCACTGAGATCCAAACTCATAT CAGTCTATAGATGTGGATGCCT	94-120; 35 [94-30; 58-30; 72-30] 72-360	319
K115 Full Provirus	5K115 GAG	S AS	AGCACTGAGATCCAAACTCATAT GGATCTCTYGTCGACTTGTCC	95-120; 35 [95-30; 58-30; 72-90] 72-360	1113
K103 Insertion site	5K103 3K103	S AS	CCACCATCTGAGAAGTGTGATG GGCAACAAAGGGTTCATATGAGAA	35 [94-60; 60-60; 72-60]	229
K103 Full provirus (5' end)	5K103 GAG	S AS	CCACCATCTGAGAAGTGTGATG GGATCTCTYGTCGACTTGTCC	94-120; 35 [94-20; 61-60; 68-84]	1188
K103 Full provirus (3' end)	ENV 3K103	S AS	GCAGGTKTAMCCAACAGCTC GGCAACAAAGGGTTCATATGAGAA	94-120; 35 [94-30; 59-90; 72-90] 72-360	1137
K103 Solitary LTR	5K103 3K103	S AS	CCACCATCTGAGAAGTGTGATG GGCAACAAAGGGTTCATATGAGAA	94-120; 35 [94-20; 61-60; 68-84]	1198
K106 Insertion site	5K106 3K106	S AS	TCCACCTGCGGACCTCCTCT TATTGGTGACAGAGAGATGCAG	94-120; 35 [94-30; 58-30; 72-30] 72-360	239

Table 2.3 ***Continued.*** Primers and Conditions used for PCR Amplification to Confirm HERV-K(HML-2) Allelic Variants. (1°) refers to primers used in primary PCR amplification and (2°) refers to nested primers used in the secondary PCR amplification.

Region Amplified	Name	Orientation	Sequence (5'-3')	Cycle (seconds)	Size (bp)
K106 Full Provirus (Hemi-nested PCR)	(1°)5K106	S	TCCACCTGCGGACCTCCTCT	94-120; 35 [94-30; 58-	
	GAG	AS	GGATCTCTYGTCGACTTGTCC	30; 72-90] 72-360	1368
	(2°)K106LS	S	TCTCTTTGGCTGGTGTGGGGA	94-120; 35 [94-30; 58-	
	GAG	AS	GGATCTCTYGTCGACTTGTCC	30; 72-90] 72-360	1229
K106 Solitary LTR (Nested PCR)	(1°)5K106	S	TCCACCTGCGGACCTCCTCT	95-120; 35 [94-30; 58-	
	3K106	AS	GGCAACAAAGGGTTCATATGAGAA	30; 72-90] 72-360	1198
	(2°)K106LS	S	TCTCTTTGGCTGGTGTGGGGA	95-120; 35 [94-30; 58-	
	K106LAS	AS	ATTCCACCAGCCTGTAGGGGA	30; 72-90] 72-360	899
K107 Insertion site	5K107	S	GGACACCCAACCTGCATGGT	94-120; 35 [94-60; 58-	646
	3K107	AS	ACACCACTGACAGTTACAGTACC	60; 72-60]	
K107 Full Provirus (5' end)	K107S	S	TCAACTCACTGCTGTGGGGAA	94-120; 35 [94-30; 58-	1230
	GAG	AS	GGATCTCTYGTCGACTTGTCC	90; 72-90] 72-360	
K107 Full Provirus (3' end)	ENV	S	GCAGGTKTAMCCAACAGCTC	94-120; 35 [94-30; 59-	1497
	3K107	AS	ACACCACTGACAGTTACAGTACC	90; 72-90] 72-360	
K107 Solitary LTR (Hemi-nested PCR)	(1°)5K107	S	GGACACCCAACCTGCATGGT	94-120; 35 [94-30; 58-	1088
	K107AS	AS	GCCGGAGGTTGTGTAGGGG	90; 72-90] 72-360	
	(2°)K107S	S	TCAACTCACTGCTGTGGGGAA	94-120; 35 [94-30; 58-	970
	K107AS	AS	GCCGGAGGTTGTGTAGGGG	90; 72-90] 72-360	

Table 2.3 ***Continued.*** Primers and Conditions used for PCR Amplification to Confirm HERV-K(HML-2) Allelic Variants. (1°) refers to primers used in primary PCR amplification and (2°) refers to nested primers used in the secondary PCR amplification.

Region Amplified	Name	Orientation	Sequence (5'-3')	Cycle (seconds)	Size (bp)
K108 Insertion Site	5K108 3K108	S AS	GTTACAGGAGTGCGCCATCAC GAATTAGGCTTTCGGGACTTCA	94-120; 35 [94-60; 58-60; 72-60]	265
K108 Full Provirus (5' end)	5K108 GAG	S AS	GTTACAGGAGTGCGCCATCAC GGATCTCTYGTCGACTTGTCC	94-120; 35 [94-30; 58-30; 72-90] 72-360	1121
K108 Full Provirus (3' end)	ENV 3K108	A AS	GCAGGTKTAMCCAACAGCTC GAATTAGGCTTTCGGGACTTCA	94-120; 35 [94-30; 59-90; 72-90] 72-360	1185
K108 Tandem Repeat	GAG ENV	S AS	GGATCTCTYGTCGACTTGTCC GCAGGTKTAMCCAACAGCTC	94-120; 35 [94-30; 58-30; 72-90] 72-360	1141
K108 Solitary LTR (Nested PCR)	(1°)5K108 3K108 (2°)K108LS K108LAS	S AS S AS	GTTACAGGAGTGCGCCATCAC GAATTAGGCTTTCGGGACTTCA AGAGATGGGTTTCTGTGGGGA GATGGTGGAAACCTGTAGGGGG	94-120; 35 [94-30; 58-30; 72-90] 72-360 94-120; 35 [94-30; 58-30; 72-90] 72-360	1162 960
3q27.2 Proviral LTR	AC69420S PROKLTR2	S AS	TGAGACAGGTACATGTGGGGAA CAGTCTATAGATGTGGATGCCT	94-120; 35 [94-30; 58-30; 72-30] 72-360	278
3q27.2 Full Provirus	AC69420S GAG	S AS	TGAGACAGGTACATGTGGGGAA GGATCTCTYGTCGACTTGTCC	94-120; 35 [94-30; 58-30; 72-90] 72-360	1229
3q27.2 Solitary LTR	AC69420S AC69420AS	S AS	TGAGACAGGTACATGTGGGGAA GTATTTTATGTTATGTACCTGTAGG	94-120; 35 [94-30; 58-30; 72-90] 72-360	960

Table 2.4 Primers and Conditions used for PCR Amplification and Sequencing of HERV-K(HML-2) Pre – Integration Sites and Flanking Sequences. (1°) refers to primers used in primary PCR amplification and (2°) refers to nested primers used in the secondary PCR amplification.

Region Amplified	Name	Orientation	Sequence (5'-3')	Cycle (seconds)	Size (bp)
7q21.2 LTR	(1°)57p21S1	S	CCACTGTGTACAAGTATATGTG	94-120; 35 [94-30; 50-	
	PROKLTR	AS	AATGGAGTCTCCYATGTCTACT	30; 72-60] 72-360	569
	(2°)57p21S2	S	GAGTCAGGGTCTCTTCTGTTG	94-120; 35 [94-30; 50-	
	PROKLTR	AS	AATGGAGTCTCCYATGTCTACT	30; 72-60] 72-360	441
7p21.2 Insertion Site (Hemi-nested PCR)	(1°)57p21S1	S	CCACTGTGTACAAGTATATGTG	94-120; 35 [94-30; 50-	401
	37p21	AS	GATTGCTCTTATAAGTCAGTTTGA	30; 72-90] 72-360	
	(2°)57p21S2	S	GAGTCAGGGTCTCTTCTGTTG	94-120; 35 [94-30; 50-	
	37p21	AS	GATTGCTCTTATAAGTCAGTTTGA	30; 72-90] 72-360	273
17q22 Insertion Site (Hemi-nested PCR)	(1°)517q22S1	S	GATTGCTCTTATAAGTCAGTTTGA	94-120; 35 [94-30; 50-	
	317q22	AS	GGGTGCAGCACACCAACATG	30; 72-90] 72-360	442
	(2°)517q22S2	S	GGGATCTTACAGATACACCAGT	94-120; 35 [94-30; 50-	
	317q22	AS	GGGTGCAGCACACCAACATG	30; 72-90] 72-360	191
17q22 LTR	(1°)517q22S1	S	GATTGCTCTTATAAGTCAGTTTGA	94-120; 35 [94-30; 50-	
	PROKLTR	AS	AATGGAGTCTCCYATGTCTACT	30; 72-90] 72-360	182
	(2°)517q22S2	S	GGGATCTTACAGATACACCAGT	94-120; 35 [94-30; 50-	
	PROKLTR	AS	AATGGAGTCTCCYATGTCTACT	30; 72-90] 72-360	157
3p25 5'Flank and LTR	53p25	S	CCGAGCTCTGTTGCACATGA	94-120; 35 [94-30; 58-	404
	GCLTRAS	AS	CATGCTGCCTTCAAGCATCTG	30; 72-30] 72-360	

Table 2.4 ***Continued.*** Primers and Conditions used for PCR Amplification and Sequencing of HERV-K(HML-2) Pre – Integration Sites and Flanking Sequences.

Region Amplified	Name	Orientation	Sequence (5'-3')	Cycle (seconds)	Size (bp)
3p25 3'Flank and LTR	GCLTRS 33p25	S AS	CTAAGGGAACTCAGAGGCTG CAGACTAAGACGTATGACTGC	94-120; 35 [94-30; 58-30; 72-30] 72-360	338
19p13.11b 5'Flank and LTR	519p13 GCLTRAS	S AS	GTACTCATACACACTGACCAG CATGCTGCCTTCAAGCATCTG	94-120; 35 [94-30; 58-30; 72-30] 72-360	453
19p13.11b 3'Flank and LTR	GCLTRS 319p13	S AS	CTAAGGGAACTCAGAGGCTG CTGGGTGTAGTCTGACTGAGT	94-120; 35 [94-30; 58-30; 72-30] 72-360	328
Xq13.1 LTR	5Xq13.1 PROKLTR2	S AS	GCATTCTCTCACAATTATGCTAC CAGTCTATAGATGTGGATGCCT	94-120; 35 [94-30; 50-30; 72-60] 72-360	479
Xq13.1 Solitary LTR	5Xq13.1 3Xq13.1	S AS	GCATTCTCTCACAATTATGCTAC GCCTGAGACTGAATGAGAGCA	94-120; 35 [94-30; 58-30; 72-90] 72-360	1314

2.3.3 Analysis of PCR Amplification Products

For analysis of amplified DNA, 20 µl of each sample was fractionated on a 2 % agarose gel (w/v) in 1 x TBE buffer (v/v) stained with 0.05 µg/ml ethidium bromide (EtBr). Depending upon the expected amplification product lengths, the gel was run for 30 to 60 mins at 150 Volts. Amplified DNA was detected due to the presence of EtBr which intercalates into double stranded DNA and exhibits fluorescence under UV light. Product size was confirmed by comparison to a 100 bp or 1 kb DNA ladder (Promega).

2.3.4 Automated DNA Sequencing

Nucleotide sequences of PCR products were determined by ABI PRISM cycle sequencing, utilising BigDye fluorescently labelled 2',3'-dideoxynucleotides (ddNTPs) (Applied Biosystems). During the strand elongation stages of the cycle sequencing reactions, DNA polymerase either incorporates standard dNTPs or analogous ddNTPs. Incorporation of a ddNTP onto the 3' end of a growing chain terminates strand elongation - selectively at G, A, T or C - as the ddNTP lacks a 3'-hydoxyl group, preventing the formation of a phosphodiester bridge with the incoming dNTP (Sanger et al., 1977).

Terminating ddNTPs are 3' labelled with four different fluorescent dyes (BigDye) which are used to identify G, A, T or C terminating reactions, due to their

property of emitting different wavelengths of light when excited by a laser. Thus all four colours and consequently all four nucleotide bases can be identified from a single cycle sequencing reaction during capillary electrophoresis.

Cycle sequencing reactions were performed using the BigDye Terminator v1.1 Cycle Sequencing Kit (Applied Biosystems) following the manufacturers instructions. Reactions were carried out in a volume of 10 μl containing; 3 - 10 ng of PCR product, 0.3 pM of either sense or antisense primer, 2 μl of BigDye v1.1 and 3 μl of the sequencing buffer supplied with the kit (complete lists of primer sequences and annealing positions used for sequencing are shown in Tables 2.1 to 2.4). Reactions were covered with a drop of liquid paraffin oil and transferred to a Techne Genius thermal cycler. Template denaturation, primer annealing and strand elongation conditions were as follows; 25 cycles of [95 ^{o}C for 20 s; 50 ^{o}C for 15 s and 60 ^{o}C for 60 s].

Prior to analysis the cycle sequencing products were ethanol precipitated, in order to remove unincorporated dye terminators. The reaction mixture was removed from below the paraffin oil and added to a microcentrifuge tube containing 50 μl of 100 % ethanol (v/v) and 2 μl of 3 M sodium acetate (pH 4.6). The mixture was mixed by vortexing and left at room temperature for 15 min to precipitate the extension products. The tubes were then centrifuged at 15000 rpm for 20 min and the supernatant aspirated. The DNA pellet was further washed by the addition of 70 % ethanol (v/v) followed by centrifugation at 15000 rpm for 5 min. The supernatant was again aspirated and the DNA pellet dried on a hot block at 55 ^{o}C for approximately 10 min. The labelled extension products were then sent to a DNA

sequencing service, in pellet form, where they were analysed on and ABI PRISM sequencing machine by capillary electrophoresis and excitation with an argon laser.

Sequencing results were viewed in the CHROMAS sequence viewer and directly imported into the SIMMONICS package for alignment and further analysis.

2.4 Population Genetic Analysis

Population genetic analysis of the variable HERV-K(HML-2) loci discussed in Chapter 4 involved the utilisation of the programmes 'Genetic Analysis in Excel' (GenAIEx) version 5.04 (Peakall and Smouse, 2001) and 'PopGene' version 3.1 (Yeh et al., 1997). Using the GenAIEX programme an Excel spreadsheet was generated by selecting 'Create a File' with the option of 'Co-Dominant Data'. Co-Dominant was selected as HERVs are regarded to be selectively neutral. The Data parameters of; 4 loci, 109 samples (the total of the population sizes of 25, 28, 22 and 34), 4 populations and two alleles were entered.

Following the generation of a spreadsheet, cells containing the designation of population number were changed to the name of the respective population. Information regarding the genotype was then entered for each of the samples examined. Here, two columns were presented for each sample as there were two possible alleles. Heterozygous individuals were scored as (1; 1) and homozygotes (0; 2) or (2; 0) depending upon their genotype. This spreadsheet was saved as an ASCII file and exported to PopGene using the GenAIX drop down menu.

Following the importation of the ASCII file into PopGene the drop down menu 'Co-Dominant' was selected and the option of 'Diploid Data' chosen. For the diploid data analysis, all options were selected. For the 'Hierarchal Structure' of the data this included the options to examine it as; single populations, groups and multiple populations. For the 'Single locus' tests all options were again selected and included; genotype frequency, allele frequency, Hardy-Weinberg tests, observed and expected homozygosity and heterozygosity, F-Statistics, gene flow and genetic distance. Following the output of all the test results, each was analysed as described in Section 4.2.5.

2.5 Analysis of Sequence Data

2.5.1 Construction of Neighbour - Joining Trees

After HERV sequence datasets were generated and aligned by hand, phylogenetic trees were constructed using the neighbour-joining method. Each tree was constructed using MEGA version 2.1 (Kumar et al., 2001), using the Kimura-2-parameter distance estimate and the option of handling gaps as a complete deletion.

The neighbour-joining method constructs a minimum evolution tree by sequentially finding pairs of extant sequences which are connected by only a single interior node. This method does not attempt to cluster the most closely related extant sequences, it minimises the length of all internal branches and so the length of the entire tree (Saitou and Nei, 1987).

Generally, when calculating distance estimates between sequences, it is expected that the transversion of a nucleotide is more common than the transition as only one of the three possible substitutions of a nucleotide is a transition. However, it has been observed that substitutions in the form of transitions arise more frequently than those resulting in transversions. When comparing extant sequence data, the Kimura-2-parameter accounts for the unequal rate at which transitions and transversions have been shown to accumulate (Kimura, 1980).

Bootstrap analysis was conducted to evaluate the reliability of inferred neighbour-joining trees, using the function available in MEGA-2. In this method a new bootstrap replicate alignment is constructed by selection of random windows of the original alignment. This process is continued until a new alignment which is the same length as the original is constructed. Each window can be selected any number or times or not at all; this leads to an alignment in which some of the original sequence characters are duplicated and others are omitted. A phylogenetic tree of the bootstrap alignment is then constructed. This process is then repeated and can result in 100 to 1000 bootstrap trees. Following this, the proportion of each clade among all bootstrap replicates is then calculated, which is expressed as a percentage confidence interval (Felsenstein, 1985).

In this study, unless stated, 500 bootstrap replicates were performed for each dataset. Branches that were supported by bootstrap confidence intervals of less than 70 % were treated with caution.

2.5.2 Construction of Maximum Parsimony Trees

After HERV sequence and related retrotransposon ORF datasets were generated and aligned by hand at the amino acid level, phylogenetic trees were constructed using the maximum parsimony method. Each tree was constructed using the program DNA-PARS as part of the PHYLIP version 3.5 package (Felsenstein, 1993) on default settings.

Using extant sequences, maximum parsimony acts to reconstruct the tree that minimizes the amount of evolutionary changes required to explain the dataset. This assumes that the most natural phylogenetic relationship of the extant data is one which requires the fewest number of changes.

Following the generation of maximum parsimony trees within the SIMMONIC sequence analysis package, the branching topology of each one was drawn out by hand.

2.5.3 Calculation of Synonymous and Non-synonymous Sequence Variability

After HERV sequence and related retrotransposon ORF datasets were generated and aligned by hand at the amino acid level, the variability at non-synonymous and synonymous sites was determined for each codon using the SIMMONIC sequence analysis package on default settings. The Jukes – Cantor method was selected to calculate sequence distances. This model assumes that the

four nucleotides have equal frequencies and that all substitutions are equally likely (Jukes and Cantor, 1969). The frequency of synonymous substitutions (dS) was divided by the frequency of non-synonymous substitutions (dN) to examine the level selection acting upon the ORF datasets.

2.5.4 Calculation of Synonymous and Non-synonymous Sequence Variability within Maximum Parsimony Trees

Following the generation of HERV-K and related retrotransposon ORF maximum parsimony trees, variability at synonymous and non-synonymous sites was determined for each codon. This was achieved by the application of the Jukes – Cantor model within the SIMMONIC sequence analysis package using the default settings. This analysis also included the examination of variability acting upon the reconstructed ancestral sequences. Following the determination of variability, all results were exported to an Excel spreadsheet for further analysis. The ratio of dS / dN was then calculated between all phylogenetically adjacent sequences. This included between extant sequences and those reconstructed by maximum parsimony. Each ratio was then assigned to its phylogenetic position within the tree topology. Ratios positioned at terminal branches were assigned to group 'A' and the ratios present at internal branches were assigned to groups 'B' or 'C'. Ratios contained within group 'C' represented those furthest removed from the extant sequences. The mean ratio of each of the groupings was then calculated.

CHAPTER 3

SCREENING FOR HERV-K WITHIN THE HUMAN GENOME

3.1 Introduction

Human endogenous retroviruses (HERVs) have been divided into three broad classes according to their genomic sequence similarities to mammalian exogenous retroviruses. Class I HERVs show similarity to the gammaretroviruses (type C retroviruses), class II to the betaretroviruses and alpharetroviruses (type B and D retroviruses) and class III are distantly related to the spumaretroviruses. These three major groupings have been further subdivided into families on the basis of sequence similarity and putative primer binding site specificity (PBS). For example, class II HERVs are often collectively referred to as the HERV-K superfamily as they are primed by a Lys tRNA. Where PBS affinity cannot be determined, HERVs have been named according to a nearby gene, specific amino acid motif or after the author who first reported the sequence.

For several reasons this nomenclature or 'grouping' of HERV families is extremely ambiguous. Firstly, highly divergent retroviruses can share the same PBS tRNA. For example, some members of class I and II endogenous retroviruses are primed by $tRNA^{Lys}$ while others in class I and III HERVs are primed by $tRNA^{Leu}$ (Tristem, 2000). Secondly, some HERVs have also been classified into groups based upon sequence similarity within a specific genic region; which does not reflect the entire HERV sequence structure. For instance the HERV-K superfamily has been subdivided into subgroups HML-1 to HML-l0, based upon sequence divergence within the *pol* region (Andersson et al., 1999). However further analysis has revealed that each subgroup is distinct as they possess LTRs that are unrelated to each other. Furthermore detailed analysis of several of the subgroups has revealed that not all the

HERV-K subgroups are primed by $tRNA^{Lys}$ although they possess similar *pol* regions (Tristem, 2000; Jurka, 2000; Lavie et al., 2004) (Table 3.1). Finally, HERVs are highly recombingenic. This has led to the mosaic evolution of viral families which hampers the reconstruction of ancient viral genotypes (Costas and Naveira, 2000). Furthermore, novel HERV families have also been generated, such the chimeric HERV-H / HERV-K retroelement and the SVA retrotransposon family (Lapuk et al., 1999; Zhu et al., 1994; Ostertag et al., 2003).

Of the 26 distinct HERV lineages identified within the human genome (Tristem, 2000; Benit et al., 2001), the majority of elements described are defective as they contain stop-codons, frameshifts and large scale insertions and deletions within their ORFs. However, many remain transcriptionally active (Goodchild et al., 1995; Seifarth et al., 1995; Seifarth et al., 1998; Huh et al., 2003). The HERV-K(HML-2) subfamily is acknowledged to be the most biologically active as it has retained the ability to encode functional retroviral protein (Towler et al., 1998; Tonjes et al., 1999; Berkhout et al., 1999; de Parseval et al., 2003), produce retrovirus-like particles (Boller et al., 1993; Lower et al., 1993a; Simpson et al., 1996) and maintain ORFs (Zsiros et al., 1999). One member of this family (HERV-K108 / HERV-K(C7)) carries only a single inactivating point mutation within *pol* which has been shown to be polymorphic in humans (Mayer et al., 1999; Tonjes et al., 1999). It is estimated that two out of seven individuals will possess the variant which can encode an intact virus (Reus et al., 2001a).

Prior to this study, eight HERV-K(HML-2) proviruses had been identified which were unique to humans (Tonjes et al., 1999; Barbulescu et al., 1999), suggesting that this family has remained retrotranspositionally active following the

Table 3.1 HERV-K classification. [a] As named within the literature or within Repbase (Jurka, 2000). [b] Number of proviruses estimated by Southern blotting. ND – Not Determined.

HERV-K	Other Reported Names [a]	Primer tRNA	Size Provirus (bp)	Size LTR (bp)	No. Intact Proviruses	Reference
HML-1	HERVK14I, LTR14A, 14B, NMWV6	Lys	ND	ND	ND	
HML-2	HERVK, LTR5, HERK, HERV-K10, HTDV, NMWV1	Lys	9500	968	30-50 [b]	(Mueller-Lantzsch et al., 1993)
HML-3	HERVK91, MER9, HERV-L70A, HERVK76, HERV50, NMV5	Lys	7913	512	14	(Mayer and Meese, 2002)
HML-4	HERVK131, LTR13, ERV-MLN, HERV-K(T47D)	Lys	9315	950	7 [b]	(Seifarth et al., 1998)
HML-5	HERVK22I, LTR22,22A,22B, NMWV2	Met	7000	497	9	(Lavie et al., 2004)
HML-6	HERV-K31, LTR3,3B, NMWV4	Lys	6900	680	30-40 [b]	(Medstrand et al., 1997)
HML-7	HERVK11D1, MER11D, NMWV7	Lys	ND	ND	ND	
HML-8	HERVK11I, MER11A,11B,11C, NMWV3	ND	ND	ND	ND	
HML-9	NMWV9	Cys	8500	420	2	(Tristem, 2000)
HML-10	HERV-KC4, LTR14, HERV-K(C4)	Lys	6357	547	10-50 [b]	(Tassabehji et al., 1994)

evolutionary divergence of chimpanzee and human. Two insertionally polymorphic proviruses were subsequently identified (Turner et al., 2001), signifying that this family might be still active. Furthermore several HERV-K(HML-2) solitary LTRs, which are the end product of intra-element homologous recombination between the 5' and 3' LTRs of a provirus (Mager and Goodchild, 1989), have also been identified which are unique to humans (Medstrand and Mager, 1998; Lebedev et al., 2000; Kurdyukov et al., 2001; Mamedov et al., 2002; Buzdin et al., 2002; Buzdin et al., 2003).

The HERV-K(HML-2) subfamily is estimated to have initially integrated into the genomes of the primate lineage following the evolutionary divergence of New and Old World Monkeys and is composed of three proviral variants (Mayer et al., 1998). The subtype HERV-K(OLD) is believed to be the ancestral variant as it retains a *gag* region which is 96 bp longer than the two other variants (Reus et al., 2001b). The remaining two subtypes are distinguished by a 292 bp deletion within the *pol-env* boundary, with Type I elements retaining the deletion (Ono et al., 1986; Lower et al., 1993a). Both of these subtypes have remained retrotranspositionally active within the hominid lineage (Barbulescu et al., 1999; Costas, 2001; Hughes and Coffin, 2001; Macfarlane and Simmonds, 2004). The LTRs of this family are also classified into different lineages based upon diagnostic substitutions and indels (Medstrand and Mager, 1998; Lebedev et al., 2000; Reus et al., 2001b; Buzdin et al., 2003).

Before the advent of the human genome sequencing project, studies directed to the discovery of novel HERV sequences relied on low stringency screening of human genomic libraries with probes derived from conserved genic regions. As a

result, the amount of data retrieved was limited. For example, Mestrand and co-workers (Medstrand and Blomberg, 1993) determined that the human genome contained 10 to 20 HERV-K(HML-3) proviral loci using a *pol* specific oligonucleotide probe. Subsequent bioinformatic analysis by Mayer and Meese, (2002) ascertained that the human genome is comprised of approximately 140 proviral loci, fourteen of which consist of *gag*, *pol* and *env* regions flanked by two LTRs which are not interrupted by large-scale indels. The large amount of sequence data available as a result of the human genome project provides a unique opportunity for the identification, examination and reconstruction of HERV sequences.

It is estimated that the HERV-K(HML-2) subfamily is composed of 30 to 50 intact proviruses within the human genome (Mueller-Lantzsch et al., 1993) (Table 3.1). Although several studies have attempted to determine the genomic locality of these proviruses (Tonjes et al., 1999; Reus et al., 2001b; Reus et al., 2001a; Sugimoto et al., 2001; Costas 2001; Hughes and Coffin, 2001; Kurdyukov et al., 2001), none have generated a comprehensive catalogue of the total number, genomic structure, cytogenetic location and relative age of them. Likewise, a directory of the total number, genomic structure and cytogenetic location of reported human specific HERV-K(HML-2) LTRs has not been complied. Furthermore, beyond the observation that the human genome contains at least seven intact HERV-K(HML-4) proviruses and 2500 solitary LTRs (Liao et al., 1998; Seifarth et al., 1998; Seifarth et al., 1995; Baust et al., 2001), no research paper has yet examined this subfamily utilising bioinformatic resources. Each of these themes is considered within this study.

In addition, as inventories were generated of the HERV-K(HML-2), HERV-K(HML-3) and HERV-K(HML-4) proviral subfamilies, the retrotranspositional history of each is considered. The validity of the divergence of proviral LTRs in serving as a molecular clock is also investigated by comparing the estimated age of HERV-K(HML-2) proviruses to their relative age. The results imply that the LTRs of several proviruses have been subject to extensive sequence exchange following their integration into the genomes of the human lineage.

3.2 Results

3.2.1 *In Silico* Detection of HERV-K Proviruses

To identify complete and near complete HERV-K(HML-2), HERV-K(HML-3) and HERV-K(HML-4) proviral sequences within the human genome, the following accessions were used as probes to conduct BLAST searches of the redundant and non-redundant nucleotide-nucleotide sequence databases. Accession M14123 was used to detect HERV-K(HML-2) sequences (Ono et al., 1986) and AF020092 for HERV-K(HML-4) proviruses (Seifarth et al., 1998). As a consensus HERV-K(HML-3) proviral sequence was not present as a sequence submission within any of the nucleic acid sequence databases, the consensus described in Mayer and Meese, (2002) was generated by hand using the SIMMONIC sequence analysis package. Searches of different releases of the human genome project assemblies were carried out with the NCBI release 33 (14th April 2003) being the latest version used in this study (Section 2.1.2).

In order to obtain an accurate account of the total number and chromosomal location of HERV-K proviruses, a stringent protocol was followed. Sequence accessions which contained high-scoring segment pairs (HSPs) were aligned individually with 1500 bp of cellular flanking sequence also being retained. This permitted proviral sequences to be characterised according to their chromosomal location and also facilitated the analysis of proviral sequence structure. In addition, duplicate sequence accessions could be excluded and the frequency of large scale duplications (parologous sequences) examined. Orthologous proviral sequences

present in non-human primates were also characterised and retained within the dataset when available (Tables 3.2 to 3.4).

Alignment of the sequence entries belonging to the three HERV-K subfamilies was carried out following each of the published consensus sequences (Ono et al., 1986; Seifarth et al., 1998; Mayer and Meese, 2002). As many of the sequence entries appeared to be highly divergent, where putative ORFs were present the sequences were aligned at the amino acid level in order to maintain a sensible alignment. During alignment several sequence entries were observed to contain large indels. Where insertions were present within the putative ORFs, the 'insertion' was followed at the amino acid level to determine if it was part of a proviral sequence which was not present in the consensus. To further analyse the nature and origin of the insertions, each one was removed from its respective sequence entry and entered into a BLASTN search to determine its classification (Tables 3.3 to 3.5).

Throughout the duration of this study several publications were released which were also directed to determining the total number of HERV-K(HML-2) proviral sequences within the human genome (Barbulescu et al., 1999; Sugimoto et al., 2001; Tonjes et al., 1999; Costas, 2001; Kurdyukov et al., 2001; Reus et al., 2001b; Hughes and Coffin, 2001). In order to generate a composite catalogue of all proviral sequences the location and total number of sequences determined within this study was compared to those reported within this literature. Subsequently where possible, HERV-K(HML-2) sequences were assigned their bibliographic name as described the literature (Table 3.2). In total 32 proviral sequences were detected within this study with an additional 14 being recorded within the literature (Table 3.6). From the total of 46 HERV-K(HML-2) proviruses three, HERV-K 4p16

Table 3.2 HERV-K(HML-2) Proviruses contained within the Human Genome. (+) and (-), designate orientation in sequence entry. Numbers correspond to the position of the first nucleotide of the 5' LTR.

Bibliographic HERV Name	Accessions	Position in Sequence entry	Location	Proviral Type
K101	AF16409		22q11.2	I
	AC007326	36875 - 46049 (+)		
	FID 83799			
	AF164610		1q21	I
K102	AL353807	157139 - 166318 (-)		
	FID 1			
	AC044819	121721 - 130900 (-)		
K103	AF164611		10p12.1	I
	AF59796			
	AL591164	139666 - 148845 (-)		
K104	AF164612		5p14.3	II
	AC025757	3662 - 13109 (-)		
	AC116309	114121 - 123566 (-)		
K106	AF16540		3q13.2	I
	AC078785	5274 - 14432 (-)		
K107	M14123		5q33.3	I
HERV-K10	AF164613			
	AC016577	20932 - 30110 (-)		
	FID27409			
K108	AC072054	29439 - 47416 (-)	7p22.1	II
HML-2.HOM	AC0104060			
HERV-K(C7)	Y17832			
	AF164614			
	AF074086			
	FID37994			
	AF261945			
K109	AL590785	33990 - 43411 (-)	6q14.1	II
	AC055116	139321 - 148740 (+)		
K113	AY037928		19p13.11	II
K115	AY037929		8p23.1	II
	AC130464			
	AC130367	10671 – 20133 (+)		
12q14.1	AC025420	37159 – 46615 (+)	12q14.1	II
	AC074261	87074 – 96525 (+)		
	FID58908			
11q22.1	AP007776	34849 – 44289 (+)	11q22.1	II
	FID54721			
HERV-K(II)	AB047209		3q21.2	II
	AC092902			
	AC069047	121049 - 130507 (-)		
	AC092903			
	AC026957			
3q27.2	AC069420	61234 - 70443 (-)	3q27.2	I
	AC015525			
	AC133473			

Table 3.2 ***Continued.*** HERV-K(HML-2) Proviruses contained within the Human Genome. (+) and (-), designate orientation in sequence entry. Numbers correspond to the position of the first nucleotide of the 5' LTR. [a] Chimpanzee Ortholog.

Bibliographic HERV Name	Accessions	Position in Sequence entry	Location	Proviral Type
1p31.1	AC093156	159605 - 165977 (-)	1p31.1	I
4q32.3	AC106872	115025 – 122253 (+)	4q32.3	I
	AC108519			
	AC068369			
K105	AF16419		21q11.1	I
	AF260249			
	AF260253			
K110	AL121985	78059 – 87290 (+)	1q23.3	I
HERV-K18	AC068728	65676 – 74908 (+)		
	Y18890			
	FID2			
	AF164618			
	AF134984			
	AF012336			
11q23.2	AP000831	6155 - 15314 (-)	11q23.2	I
	FID54716			
10p14	AC015686	157484 – 166947 (+)	10p14	II
	FID50753			
	AL392086			
HERV-K(I)	AB047240		3q12.1	I
	AC084198	94740 - 103862 (+)		
	FID13837			
3p25	AC018829	187323 – 194215 (+)	3p25	I
	AC018809			
19p13.11a	AC011467	104668 – 118245 (-)	19p13.11	I
	AC036240	50006 - 58196 (-)		
	AC068369			
19q13.13	AC012309	19871 - 29390 (-)	19q13.13	II
6p22.1	AL121932	77077 - 85952 (+)	6p22.1	II
	AL390196			
	AL671879			
6p21.1	AL035587	47775 - 57733 (+)	6p21.1	II
4p16	AC105916	40123 – 49722 (+)	4p16	II
Xq28	AF277315	1907 - 9374 (+)	Xq28	II
	AC144385 [a]			
10q24.2	AL392107	12175 - 29462 (-)	10q24.2	I
21q21.1	AL109763	112605 – 120245 (+)	21q21.1	I
	AL163218			
	AF240627			
HERV-K(C19)	AF017229		19p12-q12	II
	AC112702	18648 - 27802 (-)		
	AC010508	1 - 6408 (+)		
	Y17833			
12q24.11	AC002350	45766 - 53761	12q24.11	II

Table 3.3 HERV-K(HML-3) Proviruses contained within the Human Genome. (+) and (-), designate orientation in sequence entry. Numbers correspond to the position of the first nucleotide of the 5' LTR. [a] Chimpanzee Ortholog. (Ins.) indicates an insertion. (ρ) indicates a deletion.

Location	Accession	Position in Sequence entry	*env* Length	Indels
1p33	AL391844	15562 - 22575 (-)	+ 96	
12q23.2	AC025577	92006 - 98929 (+)	- 96	Ins. MLT2B *gag*
5q14.3	AC117524	46653 - 56697 (+)	- 96	Δ 1076 bp *env*-LTR
4p13	AC108467	65139 - 72111 (+)	+ 96	
4q35.1	AC093824	49351 - 56249 (-)	- 96	
6q21	AC002464	80557 - 87452 (-)	+ 96	
7p13	AC073115	87479 - 94420 (-)		Δ 905 bp *env*
19p13.11	AC010615	153571 - 28678 (-)	- 96	
4q13.1	AC097648	59357 - 66509 (+)	- 96	Ins. 206 bp *env*
4q34.2	AC019163	95990 - 103037 (+)	- 96	
12q13.12	AC090058	9466 - 16502 (+)	+ 96	Δ 615 bp *env*
19q13.31	AC011455	157369 - 164038 (-)	- 96	Δ 362 bp *pol*
7q21.3	AC069292	206767 - 213818(-)	- 96	
	AC142300 [a]	4104 – 8467 (+)	- 96	

Table 3.4 HERV-K(HML-4) Proviruses contained within the Human Genome. (+) and (-), designate orientation in sequence entry. Numbers correspond to the position of the first nucleotide of the 5' LTR. (Ins.) indicates an insertion. (ρ) indicates a deletion.

Location	Accession	Position in Sequence entry	Indels
10p15.1	AL391427	53891 – 152216 (-)	Ins. 110bp *env* Δ 826bp 3' LTR
19p13.11	AC010617	19237 – 29539 (+)	
16p13.3	AC092117 AC093517	73538 – 84933 (+)	
17q21.31	AC109326		Ins. 1093bp *pol*
	AC087650	20548 – 33183 (+)	Δ 63bp *env*
8q24.3	AC139103	56527 - 68197 (-)	Ins 1240bp *gag*
	AP005976		Ins. 79bp *gag* Ins. 369bp *env*
4q13.1	AC074250	171234 - 177635 (-)	Δ *gag*
Yq11.22.1	AC007034	64310 – 70579 (+)	Δ *gag*

Table 3.5 Features of the HERV-K(HML-2) Proviruses. (Ins.) indicates an insertion. (ρ) indicates deletion. (Inv) indicates an inversion. (☾) - Stop codon. (FS) - Frameshift.

HERV	*Gag* length	Indels and coding potential of Proviral genome
K101	- 96	☾ *prt*
K102	- 96	☾ *gag*
K103	- 96	FS *gag*
K104	- 96	☾ *gag*
K106	- 96	FS *prt*
K107	- 96	☾ *env*
K108	- 96	
K109	- 96	FS *prt* Δ 29 bp *pol*
K113	- 96	
K115	- 96	
12q14.1	- 96	
11q22.1	- 96	☾ *gag*
HERV-K(II)	- 96	Δ 106 bp *gag*
3q27.2	- 96	☾ *gag*
1p31.1	- 96	☾ *prt* Δ 2846 bp *pol*
4q32.3	- 96	FS *gag* Δ 1937 bp *pol*
K105	Not applicable	Not applicable
K110	- 96	Ins 65 bp *gag*
11q23.2	- 96	FS *gag*
10p14	- 96	☾ *gag*
HERV-K(I)	- 96	FS *pol*
3p25	- 96	Δ 1937 bp *pol*
19p13.11a	- 96	Ins 6760bp 5'LTR. Δ1937 bp *pol*
19q13.13	+ 96	☾ *gag*
6p22.1	+ 96	Ins Solo LTR *pol*
6p21.1	+ 96	Ins Alu Y *gag*
4p16	Not applicable	Δ 928 bp *gag*
Xq28	- 96	Δ 2181 bp *gag–pol*
10q24.2	- 96	Ins 9598 bp 5'LTR. Δ 634 bp 5'LTR Δ1375bp *env*
21q21.1	- 96	Δ 164bp *gag* Δ 712bp 3'LTR
HERV-K(C19)	- 96	Δ 5'LTR
12q24.11	- 96	Δ 520bp *env* Δ 3'LTR

Table 3.5 *Continued.* Features of the HERV-K(HML-2) proviruses. (Ins.) indicates an insertion. (△) indicates a deletion. (Inv.) indicates an inversion.

HERV	*Gag* length	Indels and coding potential of Proviral genome
12q24.33	Not determined	
8p23	Not determined	
22q11.23	+ 96	Ins. Alu Y *env*
20q11.22	+ 96	Ins. 2 x *Alu* Y 5'LTR. Ins. Alu Y *gag* Ins. SVA *pol-env*. Δ *gag-pol*
1q23	Not determined	
9q34.13a	Not determined	
19p13.11b	Not determined	
9q34.13b	Not determined	
11q12	+ 96	Ins. *Alu* Yb8 5'LTR
8q24.3	Not determined	
7q31.3	+ 96	Not determined
16p13.3	+ 96	Δ *gag* Δ *prt* Δ *pol* Δ *env*. Inv. *gag*
1p36.21	+ 96	Δ *pol* Δ *env* Δ 3'LTR
14q11.2	+ 96	Δ 300bp 5'LTR Δ *gag* Δ *pol* Δ *env*

Table 3.6 Further HERV-K(HML-2) Proviruses contained within the Human Genome.

Bibliographic HERV Name	Accession	Proviral Type	Reference
12q24.33	AC026786		(Hughes and Coffin, 2001)
8p23	AC087342		(Hughes and Coffin, 2001)
22q11.23	AP000345	II	(Reus et al., 2001b; Hughes and Coffin, 2001)
20q11.22	AL031668	II	(Reus et al., 2001b; Hughes and Coffin, 2001)
1q23	AC015623		(Hughes and Coffin, 2001)
9q34.13a	AC449424		(IIughes and Coffin, 2001)
19p13.11b	AC078899		(Hughes and Coffin, 2001)
9q34.13b	AL136108		(Hughes and Coffin, 2001)
11q12	AC004127	II	(Reus et al., 2001b; Hughes and Coffin, 2001)
8q24.3	AF235103		(Hughes and Coffin, 2001)
7q31.3	AC004979		(Reus et al., 2001b)
16p13.3	AC004034		(Reus et al., 2001b)
1p36.21	AL023753		(Reus et al., 2001b)
14q11.2	AL136419		(Reus et al., 2001b)

(AC105916), HERV-K Xq28 (AF277315) and HERV-K 10q24.2 (AL392107), were exclusively detected within this study (Macfarlane and Simmonds, 2004).

HERV-K(HML-2) proviral sequences exist in two different types, distinguished by a 292 bp deletion at the boundary of the *pol* and *env* regions. Of the total presented here, 16 belong to the Type I category and carry the deletion, 19 are of the Type II form and 11 remain to be determined (Tables 3.2 and 3.6) (Appendix A, Alignment A.1). Notably, the HERV-K110 / HERV-K18 (AC068728) provirus belongs to the Type I genotype and not the Type II form as originally reported by Ono et al., (1986). Interestingly, although HERV-K 4q23.3 (AC106872) appears to be of Type I genotype, it in fact possesses a deletion of 283 bp and contains the last 9 bp of the diagnostic Type II region (Appendix A, Alignment A.1).

HERV-K(HML-2) genomes have also been classified according to a 96 bp deletion within the *gag* ORF, with 'older' sequences being presumed to possess the longer variant. From the total of 46 HERV-K(HML-2) proviral genomes, 27 possess the shorter variant, 7 contain the longer type and 7 remain to be determined (Table 3.5). Of the remaining two proviruses, HERV-K 4p16 (AC105916) contains a large deletion within the *gag* region which covers the range of this diagnostic region and a complete proviral sequence of HERV-K105 is not present as a sequence submission within any of the sequence databases (Appendix A, Alignment A.1).

Other than a 65 bp insertion present within the putative *gag* ORF of the HERV-K(HML-2)-K110 provirus (AL121985) (Table 3.5), all insertions appeared to be unrelated to the HERV-K genome. Several of these belonged to other retroelement families such as Alu (Table 3.5). Notably, HERV-K 6p22.1

(AL121932) contained an additional HERV-K(HML-2) solitary LTR within the *pol* region.

Sixteen HERV-K(HML-2) proviruses contained large scale deletions, of which four, HERV-K 10q24.2 (AL392107), HERV-K 21q21.1 (AL109763), HERV-K(C19) (AC112702) and HERV-K 12q24.11 (AC002350), had lost all or most of one of their LTRs (Table 3.5). Interestingly, HERV-K 4q23.3 (AC106872), HERV-K 3p25 (AC018829) and HERV-K 19p13.11a (AC011467) all shared exactly the same 1937 bp deletion within the *pol* ORF (Appendix A, Alignment A.1).

Under the presumption that large scale indels would disrupt the coding potential of the HERV-K(HML-2) proviruses, it appeared that of the 31 elements examined in this study, 15 might retain the ability to produce protein. Detailed examination of their ORFs indicated that 12 contained frameshifts or stop codons and so would not be expected to be able to produce protein (Table 3.5). The remaining 3 proviral sequences, HERV-K113 (AY037928), HERV-K115 (AC130464) and HERV-K 12q14.1 (AC025420) did not appear to possess any disruptions within their ORFs (Appendix A, Alignment A.1).

Determination of intact HERV-K(HML-3) proviral genomes within the human genome indicated that there were at least 13 near complete sequences (Table 3.3). Each of these is also reported in Mayer and Meese, (2002). Screening of the NCBI BLASTN database within this study also detected a chimpanzee ortholog of the HERV-K(HML-2) provirus located at 7q21.3 (AC069292) within the human genome. Comparison of the two proviruses and their flanking regions suggested that they were nearly identical and as the chimpanzee accession was located on

chimpanzee chromosome 7, the two shared a common origin (Figure 3.1) (Appendix A, Alignment A.4).

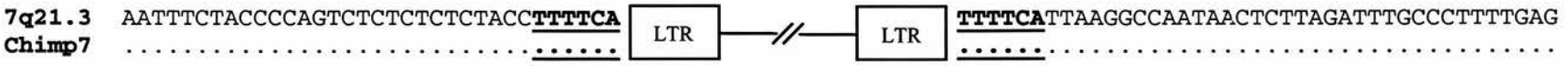
HERV-K(HML-3) proviral genomes have been classified according to the presence of 96 bp, 206 bp and 661 bp deletions within the *env* region, with older sequences being presumed to possess the longer variants (Mayer and Meese, 2002). From the total of 13 proviruses examined, 3 possessed *env* regions which were longer by 96 bp, 9 contained the deletion of 96 bp and the remaining sequence, HERV-K(HML-3) 7p13 (AC073115), possessed a large scale deletion which covered this diagnostic region (Table 3.3) (Appendix A, Alignment A.4). The HERV-K(HML-3) provirus located at 4q13.1 (AC010615) was the only provirus within the data set to possess an *env* region which was longer than 206 bp. Interestingly this insertion was not observed to be present within this provirus during the analysis by Mayer and Meese, (2002). Finally, all the proviruses examined contained a deletion of 661 bp within the *env* region when compared to the consensus (Appendix A, Alignment A.4).

Several of the HERV-K(HML-3) proviral sequences contained large scale indels. Notably, HERV-K 12q23.2 (AC025577) contained a MLT2B element within the putative *gag* region. Four of the 13 proviruses, HERV-K 5q14.3 (AC117524), HERV-K 7p13 (AC073115), 12q13.12 (AC090058) and 19q13.31 (AC011455) all possessed large scale deletions (Table 3.3) (Appendix A, Alignment A.4). Of the remaining 9 proviral sequences, all contained either frameshift or stop codons within their *gag* ORFs, indicating that they are unlikely to produce protein (Appendix A, Alignment A.6).

Analysis of proviruses belonging to the HERV-K(HML-4) subgroup

Figure 3.1 Alignment of the Flanking Regions of the Orthologous HERV-K(HML-3) Proviruses located on chromosome 7.

The direct repeats are highlighted in bold and underlined.



indicated that the human genome contains at least 7 highly divergent proviral sequences (Table 3.4). The HERV-K(HML-4) sequence located at 10p15.1 (AL391427) within the human genome shared the greatest homology to the HERV-K-T47D / HERV-KMLN provirus (AF020092) which was originally isolated from T47D particles. This included exactly the same truncation of the 3'LTR (Appendix A, Alignment A.7). As the HERV-K T47D sequence was originally identified to be located upon chromosome 10 via southern blot (Seifarth et al., 1998), it is very likely that the proviral sequence located in clone AL391427 is the same sequence.

Of the remaining 6 near intact HERV-K(HML-4) proviral sequences which were homologous to the HERV-KT47D provirus, all appeared to possess intact LTRs. However, all of the LTRs were at least 57 bp longer than those of the HERV-K T47D provirus, indicating a potentially diagnostic region approximately 400 bp upstream of the beginning of the LTR sequences (Appendix A, Alignment A.8).

Comparison of the provial ORFs indicated that the provirus located at 10p15.1 possessed an *env* region which was 110 bp longer than the rest. In addition, all of the six 'novel' proviruses possessed a longer *gag* ORF which was 221 bp longer than the transcript identified within T47D particles (Appendix A, Alignment A.7).

With the exclusion of the HERV-K(HML-4) proviruses located at 19p13.11 (AC010617) and 16p13.3 (AC092117) within the human genome, all of the identified elements contained large scale indels (Table 3.4). All proviral sequences also contained either stop codons or frameshifts within their putative *gag* region (Appendix A, Alignment A.9).

3.2.2 Catalogue of HERV-K(HML-2) Solitary LTRs

In order to facilitate the examination of the nature and spread of HERV-K(HML-2) sequences throughout the human genome, a sequence based catalogue of all HERV-K(HML-2) LTRs that were reported to date, was compiled. This allowed sequences which were duplicated within the literature to be excluded and the sequence structure of the HERV-K(HML-2) LTRs to be determined.

Prior to and during the course of this study several publications were released which were directed to discovering novel HERV-K(HML-2) LTR sequences within the human genome (Medstrand and Mager, 1998; Lebedev et al., 2000; Kurdyukov et al., 2001; Buzdin et al., 2002; Mamedov et al., 2002; Buzdin et al., 2003). Each of the reported elements was determined within this study either through the reported sequence accession or by the primers utilised within the respective study. Once a HERV-K(HML-2) LTR sequence was distinguished within a sequence accession, 1500 bp of cellular flanking sequence was retained in order to assign chromosomal location and to facilitate the exclusion of duplicate elements. Following the determination of the total number of HERV-K(HML-2) LTRs, the sequence structure of all elements was examined (Macfarlane and Simmonds, 2004).

In total, 70 HERV-K(HML-2) solitary LTRs were found within the literature (Appendix A, Alignment A.3). According to the relative age, as determined by the presence or absence of the element in non-human primates, 56 were unique to humans (Table 3.7), 9 were also present within chimpanzee (Table 3.8) and a further 5 were present in humans, chimpanzees and gorillas (Table 3.9). Interestingly, no

Table 3.7 HERV-K(HML-2) Human specific Solitary LTRs contained within the Human Genome. (△) - Deletion.

LOCATION IN HUMAN	ACCESSION NO.	SUBTYPE	FEATURES	REFERENCE
1p22.1	AF370125 / AL139421	LTR II-L / HS-a		(Buzdin et al., 2002)
1p31.2	AL356736 / AL359701	LTR II-L / HS-a		(Buzdin et al., 2002; Buzdin et al., 2003)
1q22	AL135927	Not determined		(Buzdin et al., 2003)
2p22.2	AC007390	Not determined		(Buzdin et al., 2003)
2p23.14	AC021294	LTR II-L / HS-a		(Buzdin et al., 2002)
2p23.3	AC074117	LTR II-L / HS-a		(Buzdin et al., 2002; Buzdin et al., 2003)
2q21.2	AC084028 / AC093787	LTR II-L / HS-a		(Buzdin et al., 2002)
2q33.2	AC074019	LTR II-T	Δ LTR	(Mamedov et al., 2002)
3p12.3	AF042089	LTR II-L		(Buzdin et al., 2002)
3p21.31a	Z84493 / AL450422	HS-b		(Medstrand and Mager, 1998)
3p21.31b	AC025548 / AC104447	Not determined		(Buzdin et al., 2003)
3q26.31	AC068566 / AC104640	LTR II-L / HS-b		(Buzdin et al., 2002)
3q28	AC062008 / AC112909	LTR II-L		(Buzdin et al., 2002)
4q13.3	AC055844 / AC106051	HS-b		(Buzdin et al., 2003)
5p15.31	AC091985	LTR II-L4		(Mamedov et al., 2002)
5q23.1	AC010267 / AC108095	LTR II-L / HS-b		(Buzdin et al., 2002)
5q35.1	AC008648	Not determined		(Buzdin et al., 2003)
5q35.3	AC023559 / AC113425	LTR II-L / HS-a		(Buzdin et al., 2002)
6q15	AL021774	LTR II-L / HS-a		(Buzdin et al., 2002)
6q23.2	AL596188	LTR II-L4		(Mamedov et al., 2002)
6p21.32a	Z80898 / U92032	HS-b		(Medstrand and Mager, 1998)
6p21.32b	AC022567 / X87344	LTR II-T		(Buzdin et al., 2002)
7p21.2	AC006035	LTR II-L4		(Mamedov et al., 2002)

Table 3.7 ***Continued*** HERV-K(HML-2) Human specific Solitary LTRs contained within the Human Genome. (△) - Deletion.

LOCATION IN HUMAN	ACCESSION NO.	SUBTYPE	FEATURES	REFERENCE
7q31	AC006029	LTR II-L / HS-a		(Buzdin et al., 2002)
7q31.3	AC02508	LTR II-L / HS-a		(Medstrand and Mager, 1998; Lebedev et al., 2000)
7q31.33	AC019155	LTR II-L4		(Mamedov et al., 2002)
9q22.2	AC015640	Not determined		(Buzdin et al., 2003)
9q12	AL39220 / AL773545	LTR II-L / HS-a		(Buzdin et al., 2002)
9q21.12	AL162412	LTR II-L / HS-a		(Buzdin et al., 2002)
9q33.2	AL359644	LTR II-L4		(Mamedov et al., 2002)
9q34.13	AL158039 / AL354855	LTR II-L / HS-a		(Buzdin et al., 2002)
11p15.4	AC018539 / AC080023	LTR II-L4		(Mamedov et al., 2002)
11q12.3a	U73641 / AP001591	HS-a		(Medstrand and Mager, 1998)
11q12.3b	AC003023 / AP003306	LTR II-L		(Buzdin et al., 2002)
11q13.3	AP001184	LTR II-L / HS-a		(Buzdin et al., 2002)
11q21.31	AP002513 / AC021820	LTR II-L4	Δ LTR	(Mamedov et al., 2002)
12p11.21	AC068887 / AC048344	Not determined		(Buzdin et al., 2003)
12p13.31a	U47924	LTR II-L / HS-a		(Medstrand and Mager, 1998)
12p13.31b	AC006432	LTR II-L / HS-a		(Buzdin et al., 2003)
12q13.13	AC027750	LTR II-L / HS-a		(Medstrand and Mager, 1998; Buzdin et al., 2003)
12q13.3a	AC079034	LTR II-L / HS-a		(Buzdin et al., 2002)
12q13.3b	AC024884 / AC025574	LTR II-L / HS-a		(Buzdin et al., 2002; Buzdin et al., 2003)
14q22.2	AL352982	Not determined		(Buzdin et al., 2003)
14q23.3	AL139022	LTR II-L / HS-a	Δ LTR	(Buzdin et al., 2002)
16p12.3	AC002400 / AC008870	HS-b		(Medstrand and Mager, 1998)
16p13.12	AC009167	LTR II-L	Δ LTR	(Buzdin et al., 2002)

Table 3.7 ***Continued***. HERV-K(HML-2) Human specific Solitary LTRs contained within the Human Genome. (ρ) - Deletion.

LOCATION IN HUMAN	ACCESSION NO.	SUBTYPE	FEATURES	REFERENCE
16q23.1	AC009132	LTR II-L4	Δ LTR	(Mamedov et al., 2002)
17p13.2	AC012146	LTR II-L / HS-b		(Buzdin et al., 2002)
17q21.2	AC068014	LTR II-L		(Buzdin et al., 2002)
17q22	AC032016 / AC000389	LTR II-L / HS-b		(Buzdin et al., 2002)
19q13.31	L47334 / AC073898	LTR II-L2 / HS-b		(Medstrand and Mager, 1998; Buzdin et al., 2002; Lebedev et al., 2000)
20q11.22	AL121753	Not determined		(Buzdin et al., 2003)
21q22.3	Q39E10 / AP001631	LTR-L / HS-a		(Kurdyukov et al., 2001)
Xp22.13	AC009858 / AL732371	LTR II-L2		(Buzdin et al., 2002)
Xq21.31	AL162723	LTR II-L2		(Mamedov et al., 2002)
Xq26.3	AL359703	LTR II - L	SVA	(Buzdin et al., 2002)

Table 3.8 HERV-K(HML-2) Solitary LTRs present in Chimpanzee and Human. (ρ) - Deletion.

LOCATION IN HUMAN	ACCESSION NO.	SUBTYPE	FEATURES	REFERENCE
2	AC027778	LTR II - L	Δ LTR	(Buzdin et al., 2002)
5q23.1	AC008553 / AC108095	LTR II - L		(Buzdin et al., 2002)
6	AL157379	LTR II - L	Δ LTR	(Buzdin et al., 2002)
6q25.1	AC023201 / AL590543	HS - b		(Buzdin et al., 2003)
7q22.3	AC004840	LTR II - L		(Buzdin et al., 2002)
8q21.3	AC068510 / AC106038	LTR II - L		(Buzdin et al., 2002)
11q13.3	AC003023 / AP002793	HS - b		(Medstrand and Mager, 1998)
19p13.3	AC022148	Not determined		(Buzdin et al., 2003)
Xp22.1	AC005867 / AC093011	LTR II - L		(Buzdin et al., 2002)

Table 3.9 HERV-K(HML-2) Solitary LTRs present in Gorilla, Chimpanzee and Human

LOCATION IN HUMAN	ACCESSION NO.	SUBTYPE	FEATURES	REFERENCE
9q34.2	AL445931	Not determined		(Medstrand and Mager, 1998)
21q11.2	AL109748	Not determined		(Kurdyukov et al., 2001)
21q21.1	AP000432	Not determined		(Kurdyukov et al., 2001)
21q22.3	AL773587	Not determined		(Kurdyukov et al., 2001)
Xq13.1	AJ239320 / BX295541	Not determined		(Buzdin et al., 2002)

LTRs were reported to be present in human and gorilla which were absent in chimpanzee.

Eight of the HERV-K(HML-2) LTRs could not be distinguished from either a solitary LTR or a proviral sequence as they had lost the 5' or 3' end of their sequence (Tables 3.7 and 3.8). Interestingly, one of the human specific LTRs located at Xq26.3 (AL359703) within the human genome, appeared to be flanked at the 5' end by a sequence resembling the *env* region of a HERV-K(HML-2) provirus. In order to ascertain if this LTR was part of a truncated / near complete provirus, the LTR and its 5' flanking region was aligned against the HERV-K(HML-2) provirus K107 (M14123) (Figure 3.2). Homology to the provirus was subsequently determined to be continued for ~ 500 bp, which encompassed the start of the 3' LTR and *env* region. Further examination of the sequence upstream of this region of homology revealed a highly repetitive VNTR – like region within Xq26.3 (AL359703). As a result, the HERV-K(HML-2) LTR contained within AL359703 was then compared to the SVA_{STPA1} retrotransposon AC014162 (Ostertag et al., 2003). Comparative analysis of the two sequences indicated that the human specific HERV-K(HML-2) LTR located at Xq26.3 (AL359703) was a member of the SVA retrotransposon family and not a direct product of the retrotransposition of a HERV-K(HML-2) provirus.

Following the analysis of the structure of the LTRs, subtype as defined in the various publications was assigned (Tables 3.7 to 3.9). Within this dataset, it appears that the LTR II-L subgroup has been active throughout the diversification of the higher primates with the subtypes LTR-L4 and LTR-L2 being unique to humans.

HERV-K(HML-2) LTRs belonging to elements of greater relative age (HERV-K(OLD)) have been identified as possessing diagnostic regions / insertions

Figure 3.2 Alignment of the Human specific HERV-K(HML-2) LTR at Xq26.3 (AL359703) to the SVA_{STPA1} retroelement (AC016142) with the 3' end of the *env* region and 3'LTR of HERV-K107 (M14123) complete provirus. Nucleotide substitutions at each position are indicated with the appropriate nucleotide.

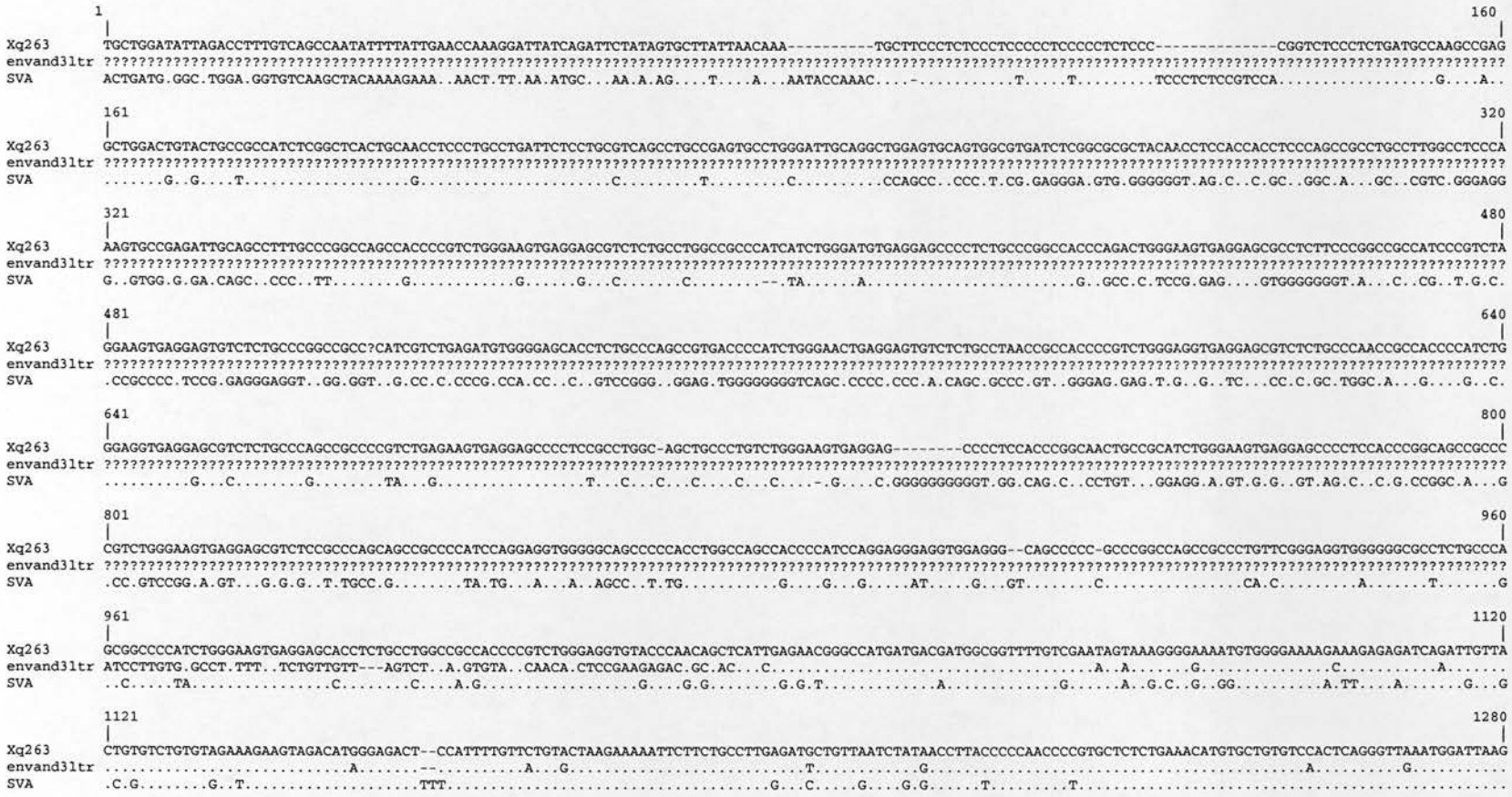
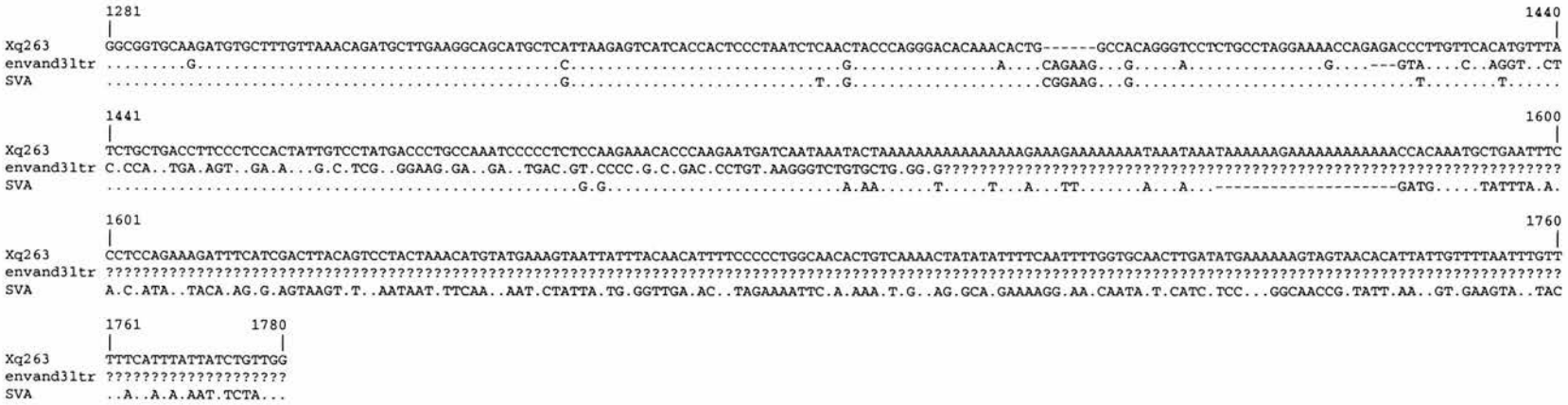


Figure 3.2 ***Continued*** Alignment of the Human specific HERV-K(HML-2) LTR at Xq26.3 (AL359703) to the SVA_{STPA1} retroelement (AC016142) with the 3' end of the *env* region and 3'LTR of HERV-K107 (M14123) complete provirus. Nucleotide substitutions at each position are indicated with the appropriate nucleotide.



of 8 bp and 23 bp which are not present in those of younger age (Mayer et al., 1998; Reus et al., 2001a). Of all the 69 HERV-K(HML-2) LTRs catalogued here, none contained these diagnostic regions, suggestive of their young age. Examination of the LTRs belonging to the 23 HERV-K(HML-2) proviruses described in Section 3.2.1, indicated that three, HERV-K 6p21.1 (AL121932), HERV-K 4p16 (AC105916) and HERV-K Xq28 (AF277315) all possessed the longer form of LTR. Interestingly, the orthologous chimpanzee sequence (AC144385) of the provirus located at Xq28 also possessed similarly longer LTRs. The presence of these two diagnostic regions within the LTRs of these three proviruses implies that they could be of considerable relative age (Appendix A, Alignment A.2).

3.2.3 Relative age of HERV-K Proviruses

During the retrotransposition of a HERV, reverse transcription generates a new retrovirus-like sequence containing two identical LTR sequences. Assuming that a provirus has not undergone any form of sequence exchange and there is no selective pressure acting on it, the accumulative nucleotide differences between the LTRs can serve as a molecular clock (Dangel et al., 1995). In order to investigate the relative age of the 46 HERV-K(HML-2), 13 HERV-K(HML-3) and 7 HERV-K(HML-4) proviral sequences described in Section 3.2.1, the number of nucleotide differences between the LTRs of each element was counted (Tables 3.10 to 3.12) (Appendix A, Alignments A.2, A.4 and A.8).

From the total of 46 HERV-K(HML-2) proviruses, HERV-K(C19) (AC112702) and HERV-K 12q24.11 (AL109763) were excluded from the data set as they had lost either their 3' or 5' LTR. In addition, the 14 HERV-K(HML-2) proviral sequences described within the literature were also not included as they were not detected within this study. This left a total of 30 HERV-K(HML-2) complete proviral sequences that could be examined (Table 3.10). All of the 13 HERV-K(HML-3) proviruses described in Section 3.2.1 were included as they all possessed fully intact LTRs (Table 3.11). Of the HERV-K(HML-4) proviral set, only the provirus located at 10p15.1 (AL391427) possessed an incomplete 3'LTR of 272bp (Table 3.12).

The number of nucleotide differences for the HERV-K(HML-2) group of proviruses varied from 0 to 78, suggestive of both recent and long term activity (Table 3.10). In contrast, members of the HERV-K(HML-3) and HERV-K(HML-4)

Table 3.10 The Relative age of HERV-K(HML-2) Proviral Sequences according to PCR Amplification and LTR Diversification.

HERV	Human	Chimp	Gorilla	Orang-utan	Gibbon	Old World Monkey	Nucleotide Differences	% Divergence between the 5' and 3' LTR	Estimated Integration date (Mya)
K101 [b]	+	-	-	-	-	-	5	0.515 (970)	0.99 - 1.981
K102 [b]	+	-	-	-	-	-	2	0.205 (971)	0.394 - 0.788
K103 [b]	+	-	-	-	-	-	6	0.625 (960)	1.201 - 2.404
K104 [b]	+	-	-	-	-	-	17	1.765 (963)	3.394 - 6.788
K106 [b]	+	-	-	-	-	-	1	0.103 (964)	0.198 - 0.396
K107 [b]	+	-	-	-	-	-	2	0.205 (973)	0.394 - 0.788
K108 [b]	+	-	-	-	-	-	6	0.617 (971)	1.186 - 2.373
K109 [b]	+	-	-	-	-	-	5	0.518 (964)	0.996 - 1.992
K113	+	-	-	-	-	-	3	0.309 (970)	0.594 -1.188
K115	+	-	-	-	-	-	14	1.441 (971)	2.771 - 5.542
12q14.1 [b]	+	-	-	-	-	-	4	0.411 (971)	0.790 - 1.581
	+	-	-	-	-	-	19	1.956 (971)	3.761 - 7.523
11q22.1 [b]	+	-	-	-	-	-	4	0.411 (971)	0.790 - 1.581
HERV-K(II)	+	-	-	-	-	-	18	2.233 (806)	4.294 - 8.588
3q27.2 [b]	+	-	-	-	-	-	3	0.308 (971)	0.592 - 1.184
1p31.1	+	-	-	-	-	-	0	0.00 (971)	0.00
21q21.1	+	-	-	-	-	-	2 (257)	0.778 (969)	1.496 - 2.972
HERV-K(C19)	+	-	-	-	-	-	Not applicable		
12q24.11	+	-	-	-	-	-	Not applicable		
4q32.3	+	+	-	-	-	-	37	3.846 (962)	7.396 - 14.792
K105	+	+	+	-	-	-	40	4.061 (985)	7.809 - 15.619
K110	+	+	+	-	-	-	34	3.49 (973)	6.711 - 13.423
11q23.2 [b]	+	+	+	-	-	-	6	0.617 (971)	1.186 - 2.373
10p14 [b]	+	+	+	-	-	-	30	2.798 (1072)	5.381 - 10.761

Table 3.10 *Continued* The Relative age of HERV-K(HML-2) Proviral Sequences.

HERV	Human	Chimp	Gorilla	Orang-utan	Gibbon	Old World Monkey	Nucleotide Differences	% Divergence between the 5' and 3' LTR	Estimated Integration date (Mya)
HERV-K(I)	+	+	+	-	-	-	18	1.842 (977)	3.542 - 7.084
12q24.33	+	+	+	-	-	-	Not determined		
8p23	+	+	+	-	-	-	Not determined		
22q11.23	+	+	+	-	-	-	Not determined		
3p25 [2]	+	+	+	+	-	-	53	5.452 (972)	10.484 - 20.969
19p13.11a [b]	+	+	+	+	-	-	62	6.378 (972)	12.265 - 24.531
20q11.22	+	+	+	+	-	-	Not determined		
1q23	+	+	+	+	-	-	Not determined		
9q34.13a	+	+	+	+	-	-	Not determined		
1p36.21	+	+	+	+	-	-	Not determined		
9q34.13b	+	+	+	+	-	-	Not determined		
19p13.11b	+	+	+	+	+	-	Not determined		
11q12	+	+	+	+	+	-	Not determined		
8q24.3	+	+	+	+	+	-	Not determined		
19q13.13	+	+	+	+	+	-	78	7.847 (994)	15.090 - 30.181
6p22.1	+	+	+	+	+	-	63 (848)	7.429 (1043)	14.286 - 28.573
6p21.1	+	+	+	+	+	+	39	3.896 (1001)	7.492 - 14.984
4p16		Not	determined				70	6.782 (1032)	13.042 - 26.084
Xq28		Not	determined				49	4.866 (1007)	9.357 - 18.715
X Chimp [a]		Not	determined				38	3.830 (992)	7.365 - 14.731
10q24.2		Not	determined				33 (335)	9.850 (1006)	18.942 - 37.884
7q31.3		Not	determined				Not determined		
16p13.3		Not	determined				Not determined		
14q11.2		Not	determined				Not determined		

Table 3.10 The Relative age of HERV-K(HML-2) Proviral Sequences according to PCR Amplification and LTR Diversification.

[a] The HERV provirus X Chimp, is an ortholog of the provirus located within the pseudoautosomal region of the human X chromosome. [b] Proviruses that were selected to determine their relative age by PCR amplification within the genomes of non-human primates. (+) indicates the presence of the HERV sequence within a primate species as determined by a positive PCR product. (-) indicates a negative PCR product and therefore the absence of the HERV sequence within the primate species. The percentage (%) divergence (D) of the LTRs of an individual element was calculated by counting the number of nucleotide differences between the LTRs and dividing this by the total length of the LTRs of that specific provirus. Where a provirus possessed a truncated LTR, the divergence was calculated using the total length of comparable sequence data. The LTR lengths are highlighted in brackets. To calculate a range of the estimated integration date of each proviral sequence, the formulae $T = D/2 \times 0.13$ and $T = D/2 \times 0.26$ were applied, where T is time passed since integration (Mya) and D is the percentage divergence of the LTRs of each specific provirus. The average mutation rates of 0.26% and 0.13% per Million years were used. This follows the estimate that endogenous retroviruses accumulate mutations within the range of 1.3×10^{-9} to 2.6×10^{-9} substitutions per site per year. The average mutation rate was multiplied by two as mutational differences could arise in either LTR of a provirus.

Table 3.11 The Relative age of HERV-K(HML-3) Proviral Sequences according to LTR Diversification. [a] Calculated as described in the legend to Table 3.10. [b] The HERV provirus 7 Chimp, is an ortholog of the provirus located at 7q21.3 within the human genome.

HERV-K(HML-3) Provirus	Nucleotide Differences	% Divergence between the 5' and 3' LTR	Estimated Integration date (Mya) [a]
1p33	27	2.192 (520)	4.215 – 8.431
12q23.2	38	7.407 (513)	14.244 - 28.477
5q14.3	44	8.577 (513)	16.494 - 32.988
4p13	31	6.25 (496)	12.019 - 24.038
4q35.1	43	8.958 (480)	17.226 - 34.454
6q21	36	7.017 (513)	13.494 - 26.988
7p13	30	6.424 (467)	12.353 - 24.707
19p13.11	45	8.772 (513)	16.869 - 33.738
4q13.1	53	10.861 (488)	20.886 - 41.773
4q34.2	53	10.351 (512)	19.905 - 39.811
12q13.12	37	7.102 (521)	12.657 - 27.315
19q13.31	35	6.849 (511)	13.171 - 26.342
7q21.3	44	8.61 (511)	16.557 - 33.115
7 Chimp [b]	40	7.827 (511)	15.051 - 30.104

Table 3.12 The Relative age of HERV-K(HML-4) Proviral Sequences according to LTR Diversification. [a] Calculated as described in the legend to Table 3.10.

HERV	Nucleotide Differences	% Divergence between the 5' and 3' LTR	Estimated Integration date (Mya) [a]
10p15.1	15	5.514 (272)	10.603 - 21.207
19p13.11	29	2.826 (1026)	5.434 - 10.869
16p13.3	26	2.541 (1023)	4.886 - 9.773
17q21.31	75	7.396 (1014)	14.114 - 28.446
8q24.3	73	7.249 (1007)	13.94 - 28.571
4q13.1	35	3.51 (997)	6.75 - 13.5
Yq11.22.1	95	9.396 (1011)	18.069 - 36.138

subgroups appeared to be less recently active with nucleotide differences ranging from 27 to 53 (Table 3.11) and 35 to 95 (Table 3.12), respectively.

Comparison of different sequence accessions corresponding to the HERV-K(HML-2) 12q14.1 provirus revealed an inconsistency. Accession AC025420, contained a total of 4 nucleotide differences whereas AC074261 comprised of a total of 19 nucleotide differences (Table 3.10) (Figure 3.3). As these two accessions could be representative of different individuals, it is possible that they represent allelic variants within the human population which may have arisen through gene conversion.

Analysis of the human and chimpanzee orthologs of the HERV-K(HML-2) Xq28 provirus, which is present within the pseudoautosomal region of the X chromosome, revealed 49 and 38 nucleotide differences respectively (Table 3.10) (Figure 3.4). Examination of the positions of the nucleotide differences revealed that 27 of them were shared by the two species and so can be presumed to have arisen within the common ancestor. Of the remaining variant sites, it appeared that 22 nucleotide differences were unique to humans and 11 were unique to chimpanzees. This suggests that since the divergence of the two species, the human lineage has accumulated 11 more nucleotide differences than the chimpanzee lineage. Interestingly 4 nucleotide sites within the LTRs were the same within each species but varied between the species, suggesting that conversion of these nucleotides occurred following species divergence.

In contrast, the LTRs of the human and chimpanzee orthologs of the HERV-K(HML-3) 7q21.3 provirus appeared to contain a comparable number of nucleotide differences (Table 3.11) (Figure 3.5). The human LTRs varied by 43 nucleotides and

Figure 3.3 Alignment of the 5' and 3' LTRs of Accessions AC025420 and AC074261 which contain the Provirus HERV-K 12q14.1.

Nucleotide substitutions at each position are indicated with the appropriate nucleotide.

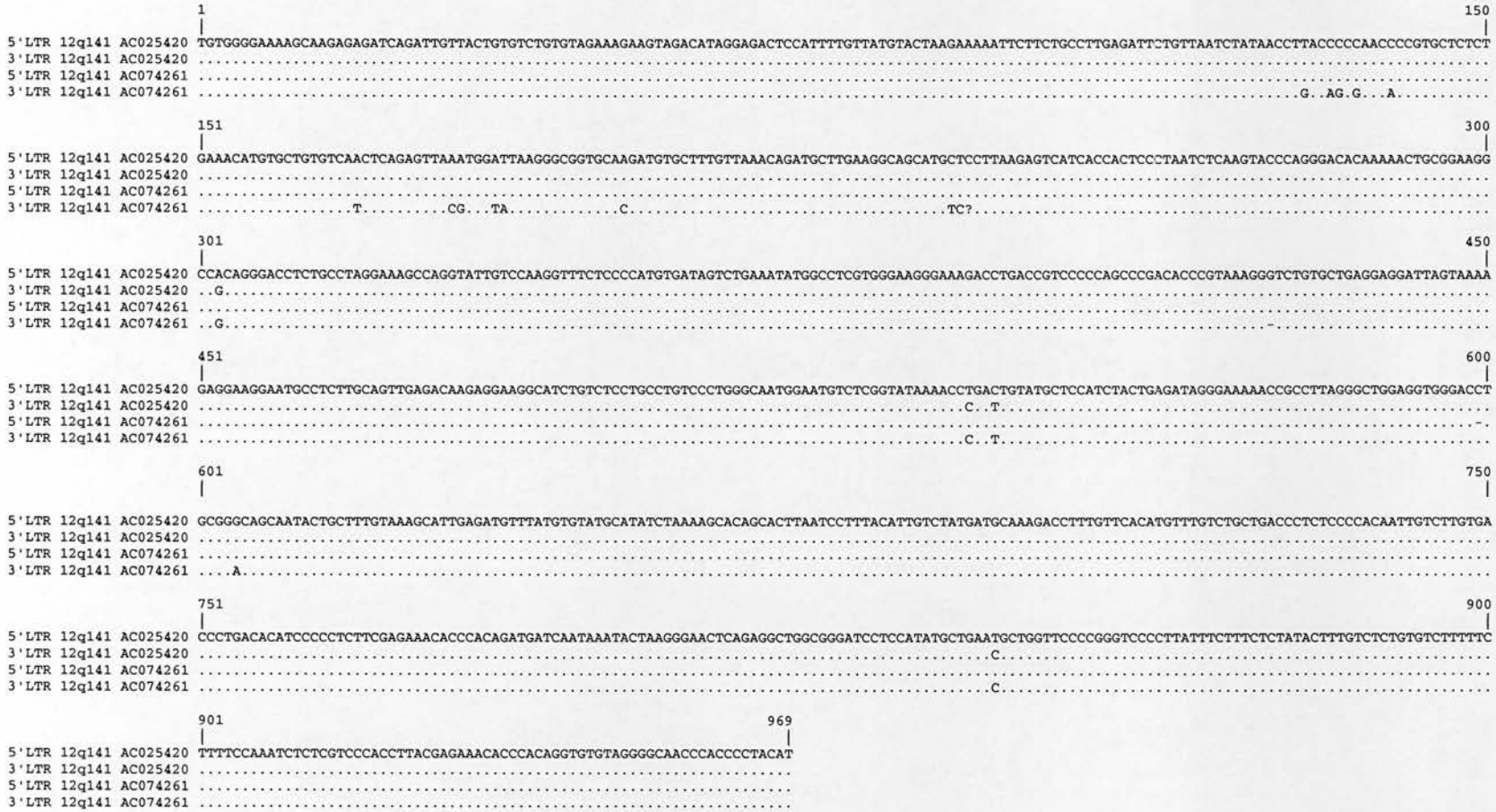


Figure 3.4 Alignment of the 5' and 3' LTRs of the Human and Chimpanzee Orthologs of the HERV-K(HML-2) Xq28 Provirus. Nucleotide substitutions at each position are indicated with the appropriate nucleotide.

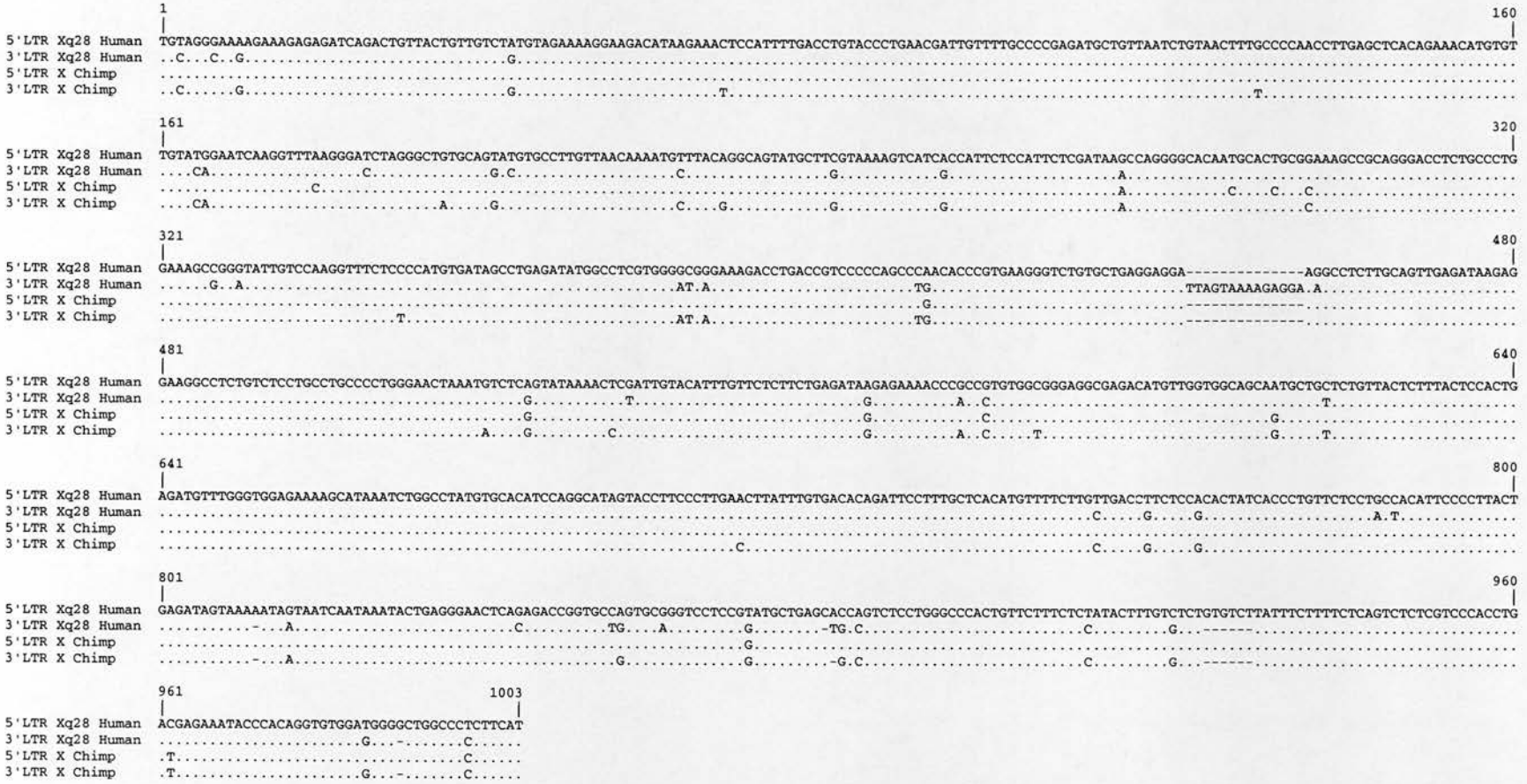
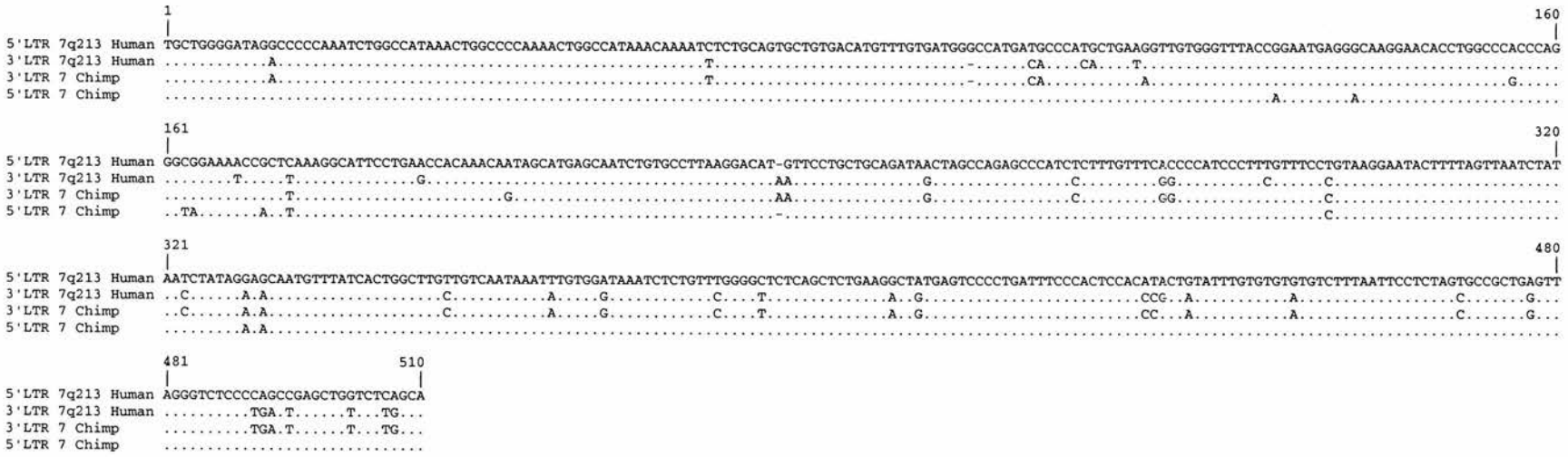


Figure 3.5 Alignment of the 5' and 3' LTRs of the Human and Chimpanzee Orthologs of the HERV-K(HML-3) 7q21.3 Provirus. Nucleotide substitutions at each position are indicated with the appropriate nucleotide.



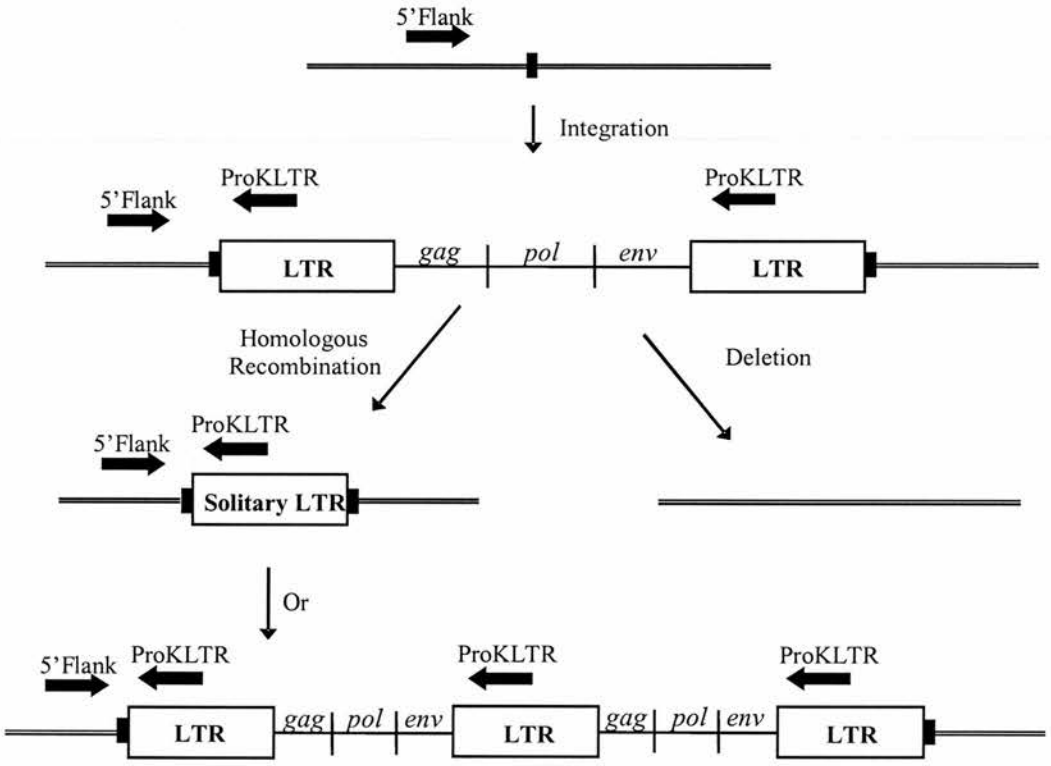
the chimpanzee by 40. Further examination of the LTRs revealed that 32 of these nucleotide differences were shared. Following the divergence of the two species, 11 unique nucleotide changes had accumulated between the human LTRs and 8 had arisen in the chimpanzee lineage.

If it is assumed that the LTRs of a provirus independently accumulate substitutions at a steady rate over time and are not subject to any from of sequence exchange, it is possible to estimate the time passed since integration via the application of a molecular clock. Within the HERV literature, estimation of mutation rates of endogenous retroviruses range from 1.3×10^{-9} to 2.6×10^{-9} substitutions per site per year (Dangel et al., 1995; Mager and Freeman, 1995; Liao et al., 1998; Johnson and Coffin, 1999; Mayer et al., 1999; Reus et al., 2001b; Lebedev et al., 2000; Turner et al., 2001). In order to calculate the relative age of each of the HERV-K proviruses listed in Tables 3.10 to 3.12, the average mutation rates of 0.13 % and 0.26 % per Mya were utilised to provide a range for the estimation of relative age. The estimated date of integration was calculated using the formula $T = D/2 \times 0.13$ (or 0.26) as described in the legend to Table 3.10.

An alternative approach to determining the relative age of HERV sequences that are present within the human genome is to amplify for their presence or absence in non-human primates. As HERVs are homoplasy-free and are unlikely to be removed from host chromosomal DNA without leaving behind a signature of their presence (Figure 3.6) (Section 4.1), it can be assumed that following their integration into germ cell DNA, subsequent host progeny will also inherit the HERV sequence.

In view of this, unique 5' flanking region primers were designed for 15 HERV-K(HML-2) proviruses according to their respective human genome sequences

Figure 3.6 Fate of HERV Proviral Sequences Following Integration. The arrows indicate the location of primers utilised for the detection of HERV-K(HML-2) proviral sequences within the genomes on non-human primates.

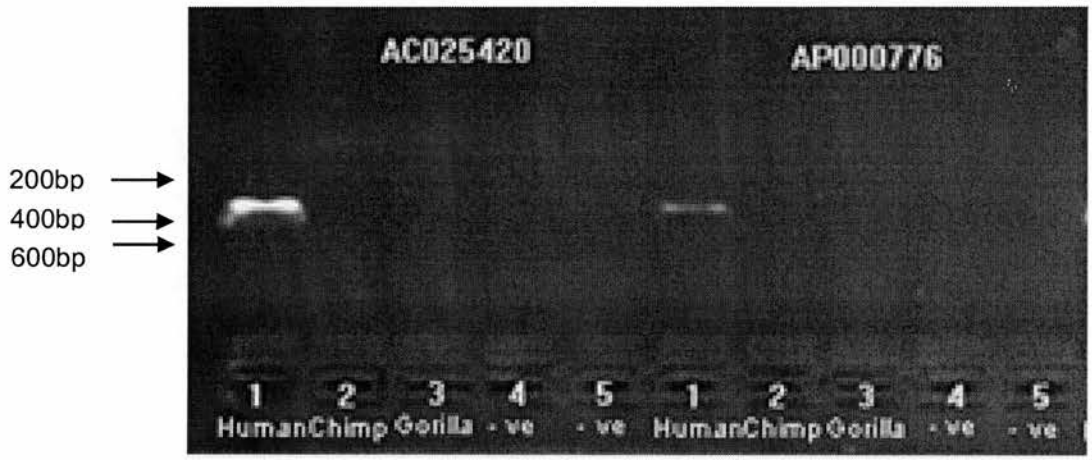


and their presence was tested for within the non-human primates Chimpanzee (*Pan troglodytes*) and Gorilla (*Gorilla gorilla*) (Table 3.10) (Figure 3.6) (all primer sequences and combinations are shown in Section 2.3.2, Table 2.2).

Prior to the PCR amplification of HERV-K(HML-2) proviruses, the quality and authenticity of the primate DNA samples was assessed (Appendix B, Figure B.2) all primer sequences and combinations are shown in Section 2.3.2, Table 2.1). This was achieved by performing a series of PCR amplification reactions. Initially, amplification was conducted for the first hypervariable segment of the primate mitochondrial genome, whereby a positive PCR product indicated that high copy number DNA was present. Following this, amplification was then conducted utilising primers specific to the human mitochondria second hypervariable region (HV2), a negative result indicated that human DNA was not present in the sample. Subsequently, amplification and sequencing of the 12S rRNA mitochondrial region was then conducted to determine the subspecies of primate. Finally, amplification for the protamine gene was performed as it provided an indication of the quality of single copy (nuclear) DNA that was present within the sample.

Of the 15 HERV-K(HML-2) proviruses that were selected for amplification, 11 appeared to be unique to humans and the remaining 4 were also present in chimpanzee and gorilla (highlighted in Table 3.10) (Figure 3.7). This implies that this HERV subgroup has remained retrotranspositionally active following the diversification of these three species. Throughout the duration of this study several publications were released which were also directed to determining the relative age of HERV-K(HML-2) proviral sequences which are present within the human genome (Barbulescu et al., 1999; Kurdyukov et al., 2001; Reus et al., 2001b; Hughes

Figure 3.7 Representative examples of the PCR determination of the relative age of HERV-K(HML-2) proviral sequences. The first set of five lanes refers to the analysis of the HERV-K(HML-2) provirus 12q14.1 (AC025420) and the second set of lanes, the HERV-K(HML-2) provirus 11q22.1 (AP007776). As both proviruses are unique to humans, they can be presumed to have integrated after the diversification of hominids and chimpanzee.

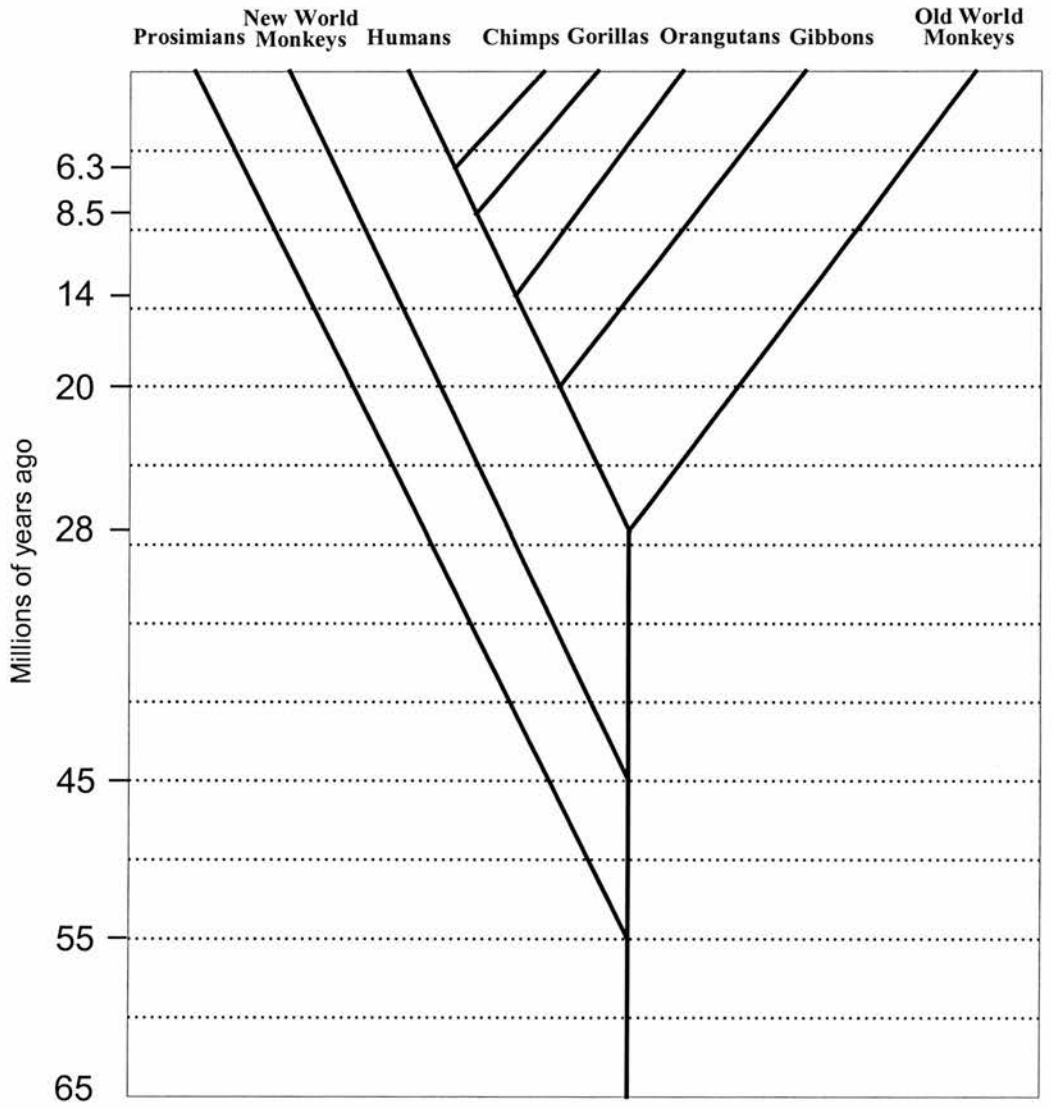


and Coffin, 2001). Where possible, each of their results is also summarised within Table 3.10.

In order to assess the validity of two LTRs of a provirus in serving as a molecular clock, the two methods of determining the relative age of each HERV-K(HML-2) provirus were compared. To assist in this comparison, primate speciation dates estimated from both archaeological and molecular data were averaged. This provided the branching dates from humans of 55 Mya for Prosimians, 45 Mya for New World Monkeys, 28 for Old World Monkeys, 20 Mya for Gibbons, 14 Mya for Orang-utans, 8.5 Mya for Gorillas and 6.3 Mya for Chimpanzee (Figure 3.8).

Using the average branching dates as a guide (Figure 3.8), of the 28 HERV-K(HML-2) proviruses that could be compared (Table 3.10), the estimated relative ages of three proviruses were not compatible. The first of these, HERV-K 11q23.2 was present in human, chimpanzee and gorilla and so is expected to have entered the genome of a common ancestor between 8.5 and 14 Mya (Figure 3.8). However, according to the divergence of the LTRs within the human ortholog (AP000831), the relative age of the provirus was between 1.186 to 2.373 Mya (Table 3.10) and so should be unique to humans (Figure 3.8). Similarly, the divergence of the LTRs of the HERV-K(I) provirus, which was present in human, chimpanzee and gorilla, also insinuated an underestimate of relative age. In this case, according to LTR divergence, the provirus entered the germline between 3.542 to 7.084 Mya (Table 3.10), implying that it was only likely to be present in human and chimpanzee (Figure 3.8). The third provirus which did not fit expectation was HERV-K 6p21.1. According to PCR amplification, it was present in a Baboon an Old World Monkey (Reus et al., 2001b), and so is expected to have entered the genome of a common

Figure 3.8 Branching Data of Primate Evolution. The branching dates were averaged from; Sarich and Wilson, (1967); Sibley and Ahlquist, (1984); Sibley and Ahlquist, (1987); Hasegawa et al., (1987); Arnason et al., (1996); Takahata and Satta, (1997); Kumar and Hedges, (1998); Stauffer et al., (2001); Chen and Li, (2001); Glazko and Nei (2003) and Schrago and Russo, (2003).



ancestor between 28 and 45 Mya (Figure 3.8). Conversely, the divergence of the LTRs implied that it entered between 7.492 to 14.984 Mya (Table 3.10). Interestingly, the relative age of this provirus was also determined by (Hughes and Coffin, 2001) who observed that it was only present in humans, chimpanzees and gorillas.

Assuming that the divergence of the LTRs of a provirus can act as a 'molecular clock' and that substitutions occur within the range of 1.3×10^{-9} to 2.6×10^{-9} per site per year, the retrotranspositional history of the HERV-K(HML-3) and HERV-K(HML-4) subgroups could also be examined. Of the 13 HERV-K(HML-3) proviruses, most members appeared to have integrated between 12 to 42 Mya (Table 3.11). This indicates that orthologs can be expected to be found in at least chimpanzee and gorilla (Figure 3.8). The exception was the provirus HERV-K 1p33, which appeared to be considerably younger with an estimated relative age of between 4.215 to 8.431 Mya. In contrast, the HERV-K(HML-4) subgroup was composed of proviruses of variable age (Table 3.12). Three, HERV-K 19p13.11, HERV-K 16p13.3 and HERV-K 4q13.1 appeared to have integrated less than 15 Mya and the remaining four, less than 37 Mya. This suggests that the HERV-K(HML-4) subgroup may have remained retrotranspositionally active during the diversification of the higher primates.

3.3 Discussion

HERV elements make up a significant proportion of the human genome (8 %) and have been proposed to be pacemakers in the evolution of primates (Sverdlov, 2000). Determining the structure and cytogenetic location of HERV sequences can be regarded as a starting point for studies investigating their impact, perhaps in regulating the expression of cellular genes or in remodelling the primate genome. In this study, 46 HERV-K(HML-2) proviruses, 62 HERV-K(HML-2) solitary LTRs, 7 incomplete HERV-K(HML-2) LTRs, 1 SVA retrotransposon, 13 HERV-K(HML-3) proviruses and 7 HERV-K(HML-4) proviruses were determined to be present within the human genome.

Prior to this study, the human genome was estimated to contain between 30 to 50 HERV-K(HML-2) proviruses (Mueller-Lantzsch et al., 1993), 14 intact HERV-K(HML-3) proviruses (Mayer and Meese, 2002) and 6 near intact HERV-K(HML-4) proviruses (Seifarth et al., 1998). The results obtained here are in keeping with these estimations. However, the cytogenetic location of the HERV-K(HML-4) proviruses determined here are not in complete accordance with those determined by Southern blot analysis. In this study, such elements were observed on chromosomes, 4, 8, 10, 16, 17, 19 and Y, whereas they were previously determined to be present on chromosomes, 10, 8, 9, 15, 16 and 19. These results may differ as the human genome contains a greater number of near intact HERV-K(HML-4) proviruses than detected by either method. Alternatively, it is possible that the Southern Blot analysis detected partial HERV-K(HML-4) sequences as the *pol* region was used exclusively as a probe (Seifarth et al., 1998). It should be considered that a large number of HERV-

K(HML-2) sequences within the human genome consist solely as *gag*, *pol* or *env* regions (Mayer et al., 1997b; Mayer et al., 1997a).

Alignment of all sequences catalogued here revealed several incongruities to reports within the literature. First, the HERV-K110 / HERV-K18 provirus appeared to be a Type I HERV-K(HML-2) provirus and not a Type II, as previously reported (Ono et al., 1986). Second, comparison of the LTRs of the HERV-K113 (HML-2) provirus revealed three nucleotide differences whereas formerly the provirus was observed to possess zero nucleotide differences (Turner et al., 2001). In addition, the HERV-K115 (HML-2) provirus was previously observed to possess a 1 bp deletion 92 bp upstream from the stop codon of the *gag* ORF which was presumed to prevent translation of the *pro* and *pol* ORFs (Turner et al., 2001). Analysis of the HERV-K115 *gag* ORF here, revealed no such deletion. Third, the HERV-K(HML-3) provirus 4q13.1 was previously reported to possess a diagnostic 206 bp deletion within the *env* ORF (Mayer and Meese, 2002), in contrast this provirus was not observed to not retain such a deletion within this study. Forth, all HERV-K(HML-4) proviruses detected here possessed both longer LTRs and a *gag* ORF than documented within the HERV-K(HML-4) consensus HERV-KT47D. Finally, the total number of human specific HERV-K(HML-2) sequences examined here was considerably less than previously reported. As all HERV-K(HML-2) LTRs sequences that had been reported to date were catalogued and examined individually within this study, duplicate accessions could be removed from the dataset. Here, the human genome contained, 18 HERV-K(HML-2) proviruses, 55 HERV-K(HML-2) LTRs and 1 SVA retrotransposon which were unique to humans. This total of 74

human specific HERV-K(HML-2) sequences is considerably less than the 130 reported elsewhere (Buzdin et al., 2002; Buzdin et al., 2003).

In combining sequence information on the genomic structure of the 46 HERV-K(HML-2) proviruses with their presence in non-human primates, the retrotranspositional history of the HERV-K(HML-2) subfamily can be examined. However, it should be considered that the genomic retroviral elements that exist today represent a small fraction of total germ line integration events, namely, those that were not detrimental to the host and that also became fixed in the genomes of the (human) primate lineage.

From the total of 46 HERV-K(HML-2) proviruses; 10 possessed the longer *gag* variant associated with older proviruses, 27 the shorter *gag* variant, 7 remain to be determined and sequence information was not available for the remaining two proviruses, HERV-K105 and HERV-K 4p16. Of the 10 'older' proviruses, the relative age of 7 has been determined via PCR amplification in non-human primates (Reus et al., 2001b; Hughes and Coffin, 2001). The oldest of these is present within members of the Cercopithecoidea and Hominoidea super families (Reus et al., 2001b) and the youngest in representatives of the Pongidae and Hominidae families (Hughes and Coffin, 2001). The results compiled here indicate that the longer *gag* variant ceased in amplification following the evolutionary split of Gorillas from the human lineage. The shorter *gag* variant appears to arisen following the divergence of the Cercopithecoidea and Hominoidea super families as is present at orthologous loci within members of the Hylobatidae, Pongidae and Hominidae families (Reus et al., 2001b; Hughes and Coffin, 2001). Amplification of this shorter variant has continued following the evolutionary divergence of humans and chimpanzees (Barbulescu et

al., 2001; Macfarlane and Simmonds, 2004). These results confirm the greater antiquity of the longer *gag* variant.

The HERV-K(HML-2) proviruses have also been classified into two types based upon a 292 bp deletion at the *pol-env* boundary, with Type I elements being presumed to be the younger variants as they carry the deletion (Ono et al., 1986; Lower et al., 1993). Here, 16 HERV-K(HML-2) proviruses were Type I, 19 were Type II, and 11 remain to be determined. Of the 7 proviruses that have a relative age determined by PCR and a longer *gag* variant; all are Type II. This association of the longer *gag* and *pol-env* regions confirms the greater age of the Type II proviral form. In accordance, the Type I proviral variant is only present at orthologous loci within all members of the Pongidae and Hominidae families so can be presumed to have arisen following the evolutionary split of Gibbons from the human lineage. This finding is in agreement with the examination of full length *env* and *gag* regions within the primate lineage (Mayer et al., 1998). Interestingly, both Type I and Type II proviral forms have remained comparably retrotranspositionally active following the evolutionary divergence of humans and chimpanzees (Macfarlane and Simmonds, 2004). Here, of the 18 HERV-K(HML-2) proviruses that are unique to humans; 8 are Type I and 10 are Type II.

Furthermore, older HERV-K(HML-2) proviral variants are classified by the presence of diagnostic 8 bp and 23 bp regions within their LTRs (Medstrand et al., 1997; Lavrentieva et al., 1998; Lebedev et al., 2000; Reus et al., 2001b). Younger variants retain these regions as deletions. Analysis of 101 HERV-K(HML-2) elements catalogued within this study revealed that only 3 possessed longer variant LTRs, each of which was a provirus. Two of these proviruses, HERV-K 4p16 and

HERV-K Xq28, were detected exclusively within this study Macfarlane and Simmonds, (2004) and the third is reported in Reus et al., (2001a) and Hughes and Coffin, (2001). The third provirus HERV-K 6p21.1 is a Type II provirus with the longer *gag* variant. HERV-K Xq28 also retains a Type II genotype but possesses the shorter form of *gag*. Finally, HERV-K 4p16 is a Type II provirus but the *gag* variant genotype cannot be determined as this provirus contains a 928 bp deletion of *gag* which encompasses this diagnostic region. These observations suggest that the longer LTR variants are associated with both the shorter or longer *gag* variants of Type II proviruses, confirming their retrotranspositional antiquity.

A further diagnostic HERV-K(HML-2) proviral region that has not been observed prior to this study is present within the *pol* ORF. Three proviruses, HERV-K 3p25, HERV-K 19p13.11a and HERV-K 4q32.3 all possessed exactly the same 1937 bp deletion within this region, signifying their relatedness. Interestingly, orthologous loci of the HERV-K 3p35 and HERV-K 19p13.11a proviruses are present within all members of the Pongidae and Hominidae families (Hughes and Coffin, 2001) whereas HERV-K 4q32.3 is assumed to have integrated within the common ancestor of human and chimpanzee (Hughes and Coffin, 2001). Each of the older proviruses is of Type I genotype, however the HERV-K 4q32.3 provirus appears to be a mosaic as it possesses a shorter deletion within the *pol-env* boundary of 283 bp and contains the last 9 bp of the diagnostic Type II region.

The validity of the divergence of LTRs of individual elements in serving as a molecular clock was considered within this study by comparing the estimated age of HERV-K(HML-2) proviruses to their relative age. To calculate an estimated range of time passed since integration, the average mutation rates of 0.13 % and 0.26 % per

Million years were used. These values are analogous to the mutation rates of, *Alu* sequences (Britten, 1994), coding and non-coding regions (Minghetti and Dugaiczyk, 1993; Chen and Li, 2001; Britten, 2002; Kumar and Subramanian, 2002) and HERVs (Liao et al., 1998; Dangel et al., 1995; Mager and Freeman, 1995; Lebedev et al., 2000) within the human genome.

Of the 27 HERV-K(HML-2) proviruses compared three, HERV-K 11q23.2, HERV-K(I) and HERV-K 6p21.1 possessed LTRs whose divergence was significantly less than expected when compared to their presence at orthologous regions in non-human primates. In these cases, if LTR divergence was used exclusively to estimate the age of the proviral insertions, the time passed since integration would be grossly underestimated.

In addition, such a method is highly sensitive to single nucleotide changes which could lead to misinterpretation of the time passed since integration. For example, according to LTR divergence, the HERV-K 4q32.3 provirus inserted 7.396 to 14.792 million years ago, in keeping with its presence in humans and chimpanzees. In contrast, the HERV-K110 provirus, which is an older insertion as it is present in human chimpanzee and gorilla, appears to be younger than HERV-K 4q32.3 as it is estimated to have inserted 6.711 to 13.423 million years ago.

Furthermore, the relative age of a provirus could be underestimated if the HERV sequence is unfixed at the time of primate species diversification. In this scenario the polymorphic variants, a pre-integration site and integrated HERV, have an equal probability of being transmitted into succeeding lineages. Hence a HERV may have integrated within a common ancestor but evidence of this insertion may not be present if the pre-integration site allele becomes fixed within a specific

lineage. Such a phenomenon has been observed within the human, chimpanzee and gorilla lineages whereby humans retain a fixed pre-integration site allele and chimpanzees and gorillas possess a HERV provirus (Barbulescu et al., 2001).

Interestingly, comparison of different sequence accessions corresponding to the human specific HERV-K 12q14.1 provirus revealed an inconsistency of estimated age, which is indicative of sequence exchange between unrelated LTRs following the integration of the provirus. If sequence exchange has occurred between related or unrelated proviral LTRs, the substitutions that have accumulated through generations of host replication will be invalidated, therefore LTR divergence will not reflect the age of a provirus. Such events are considered in more detail in Chapter 5.

However, if it is assumed that LTR divergence roughly reflects the age of a provirus the retrotranspostional history of the HERV-K(HML-3) and HERV-K(HML-4) proviruses described within this study can be examined. The HERV-K(HML-3) family appears to have been composed of two active lineages, one of which possessed a 96 bp deletion within the *env* region. Members appear to have been actively retrotransposing throughout the diversification of the Cercopithecoidea and Hominoidea super families with activity ceasing following the diversification of the Hylobatidae family. In contrast, the HERV-K(HML-4) family appears to have remained active over a longer time scale, with several proviruses potentially having integrated during the diversification of Pongidae and Hominidae families.

Of the 52 HERV-K proviral sequences identified here via screening of the human genome sequence databases, the majority appear to be incapable of producing retroviral proteins as they have acquired both point mutations and large-scale indels. However, analysis of the ORFs indicates that 3 HERV-K(HML-2) proviruses,

HERV-K113, HERV-K115 and HERV-K 12q14.1 might have retained the ability to produce retroviral protein. However, it is possible that allelic variants of the HERV-K proviruses reported within this study could exist within the human population which retain intact ORFs. Such a scenario has been reported for the HERV-K108 provirus (Reus et al., 2001a). In addition, it is also possible that insertionally polymorphic proviruses exist within the human population which are at present unidentified. This is considered in more detail within Chapter 4.

CHAPTER 4

ALLELIC VARIATION OF HERV-K

4.1 Introduction

The debate over recent human origins has primarily focused on two models. The first, referred to as the 'Multiregional model' proposes that over the last 1.5 million years, modern humans arose independently in different regions of the world but remained a single species through worldwide gene flow. In contrast, the second hypothesis which is often called the 'Out of Africa model' suggests that a single population of modern humans migrated from Africa around 100,000 to 200,000 years ago and replaced archaic human populations throughout the world. To date, the majority of studies that have examined human genetic variation conclude that the 'Out of Africa' is the more acceptable model. However, studies utilising variation contained upon the Y chromosome or mitochondrial DNA only reflect a quarter of human genetic content. Recent reanalysis of this data original data suggests that extensive admixture may have occurred between populations over a long timescale (Hammer et al., 1998; Templeton, 2002).

In examining the genetic relationships of contemporary human populations, an alternative to DNA sequence comparison would be the systematic evaluation of cladistic molecular markers such as the 'Unique Event Polymorphisms' (UEPs) of which HERVs are an example. Extensive analysis of SINE and LINE retroelements, which are the most abundant UEPs within the primate genome, has been used resolve the phylogenetic relationships of primates and human populations (Batzer et al., 1994; Boissinot et al., 2000; Schmitz et al., 2001; Roy-Engel et al., 2001; Myers et al., 2002; Ovchinnikov et al., 2002; Vincent et al., 2003; Salem et al., 2003a; Salem et al., 2003b). To date, ERVs and their derived retroelements have been used to

examine the genetic relationships of the primates (Mager and Freeman, 1995; Kim et al., 1999; Kim et al., 2000; Voisset et al., 1999; Kim et al., 2002; Huh et al., 2003; Yi et al., 2003) and to distinguish the lineage of extant and extinct Proboscidea (Greenwood et al., 2001) but none have been extensively applied to the issues surrounding the origin and dispersal of contemporary human populations.

HERV insertional or structural mutations leading to the production of a solitary LTR offer several advantages for examining human genomic diversity. First, large numbers of DNA samples can be rapidly typed using PCR-based assays. Second, as with LINE and SINE retroelements, the *de novo* insertion of a HERV sequence within the germ line represents a unique event in human genome evolution. The large number of potential target sites within the human genome and the random nature of retroviral integration denote that homoplasy is highly unlikely. Third, HERV sequences are stable as there are no known mechanisms for completely removing them without deleting host chromosomal DNA or leaving behind a solitary LTR (Johnson and Coffin, 1999). Accordingly, the directionality of the insertion and the formation of a solitary LTR can unambiguously be assigned to a specific lineage, as individual loci containing the same HERV sequence are identical by descent. Forth, the ancestral state of a HERV sequence is ultimately its absence and is represented by a pre-integration site sequence. HERV sequences that are unique to humans can be determined through PCR analysis of the orthologous region in non-human primates. This information can be used to root trees of population relationships derived from analysis of HERV polymorphisms. Finally, as the process of reverse transcription generates two LTR sequences that are identical at the time of HERV sequence insertion, the accumulative nucleotide differences between them

can serve as a molecular clock (Dangel et al., 1995). This offers an advantage over the other abundant UEPs such as the SINE and LINE retroelements as their relative age of insertion can only be predicted according to subtype or by the presence of the insertion within a specific host lineage. However, a molecular clock based upon the divergence of proviral LTRs will be invalidated if a HERV sequence has been subject to recombination or gene conversion following integration.

To ascertain the utility of HERV polymorphisms for examining recent human evolution, a total of 122 unique HERV-K insertions were screened within the human genome databases in order to determine if any of the loci were dimorphic. Seven human specific HERV-K(HML-2) loci were subsequently determined to be biallelic. Two of these were solitary LTRs which were polymorphic for insertion. The first was located at 6p21.32 and is reported to have arisen through the duplication of the MHC complex (Horton et al., 1998) and the second was located within the highly repetitive centormeric long arm of chromosome 9. The events that lead to the formation of the remaining five biallelic loci were then determined through sequence analysis which included the generation of a phylogenetic tree. The human genetic variability of seven human specific HERV-K proviral loci, which included the remaining five biallelic loci and two which were monomorphic within the human genome databases, was then determined through the PCR screening of 109 individuals from four geographically dispersed populations. Statistical population genetic analysis was then applied to determine the validity of these HERV loci as markers for examining the origin and dispersal of contemporary human populations.

4.2.1 Results

4.2.1 Screening of the Human Genome Databases

A total of 122 unique HERV-K insertions from three different subgroups and variable relative age were screened for polymorphism within the human genome databases (Table 4.1). For each insertion, 1500 bp of the 5' flanking sequence was screened by standard nucleotide-nucleotide BLAST against the non-redundant and high-throughput sequence databases, in order to detect paralogous sequences and to ascertain if polymorphism was present at specific loci. High scoring accessions that had not previously been detected during the screening for complete proviruses (Section 3.2.1) were aligned by hand in SIMMONIC against the corresponding HERV-K insertion. Cytogenetic location of the accessions was then determined according to their location within the Ensembl human genome database. Of the total, seven unique insertions appeared to be polymorphic, all of which belonged to the HERV-K(HML-2) subgroup and were human specific.

Two HERV-K(HML-2) proviruses were dimorphic for insertion with one allele representing the presence and the second the absence of the provirus. The first provirus, HERV-K113 was contained within accession AY037928 with a corresponding pre-integration site in accession AC092364. Both accessions were located in the region 19p13.11. The second provirus, HERV-K115, was contained within accession AY037929 with the pre-integration site present in accessions AF189745, AF202031, AC144950 and AC092364. All accessions corresponding to the HERV-K115 region were located at 8p23.1. To confirm that the two proviruses

Table 4.1 Total numbers of HERV-K Insertions Screened for Variance within the Human Genome Databases

Type of HERV-K element screened	Total no of unique insertion sites	Number of polymorphic loci
HML-2 Human specific provirus	18	5
HML-2 provirus present in Human and Chimp	1	0
HML-2 provirus present in Human, Chimp and Gorilla	6	0
HML-2 provirus present in Human, Chimp, Gorilla and Orang	2	0
HML-2 provirus present in Human, Chimp, Gorilla, Orang and Gibbon	2	0
HML-2 novel provirus reported in Macfarlane and Simmonds (2004)	3	0
HML-2 Human specific solitary LTRs	55	2
HML-2 solitary LTRs present in Human and Chimp	9	0
HML-2 solitary LTRs present in Human, Chimp and Gorilla	5	0
SVA Human specific retrotransposon located at Xq26.3	1	0
HML-3 provirus	13	0
HML-4 provirus	7	0

were biallelic as a result of insertion and not deletion, cellular nucleotides that are duplicated during retroviral integration were searched for by examining the regions immediately upstream and downstream of the proviruses and compared to the posited pre-integration site sequences. The provirus HERV-K113 was flanked by the direct repeat sequences CTCTAT and HERV-K115, CCTTT. Each provirus appeared to be insertionally polymorphic as the target site sequence was not duplicated in either of the pre-integration site sequences.

A further two variable loci, HERV-K103 and HERV-K106, alternated for a copy of a complete provirus and a solitary LTR. Accessions AL591164 and AC078785 contained the respective proviruses, with accessions AL139404 and AC024108 the solitary LTRs. The accessions for the HERV-K103 region both plotted to 10p12.1 and those of the HERV-K106 were located at 3q13.2. As with the insertionally polymorphic proviruses, the direct repeats of the solitary LTRs corresponded to their respective provirus.

A fifth HERV-K(HML-2) loci contained variability for a complete provirus and a tandem duplication of the provirus which consisted of the structure LTR-*gag*-*pol*-*env*-LTR-*gag*-*pol*-*env*-LTR. Accessions AC072054 and AC0104060 contained the tandem duplication, with Y17832 and AF164614, the single proviral copy. All accessions corresponded to the cytogenetic region 7p22.1 within the human genome with proviral sequences flanked both upstream and downstream by the sequence GGTTTC.

In addition to the structural and insertional HERV-K(HML-2) complete proviral variants, two solitary LTRs that are unique to humans were also identified to be insertionally polymorphic. The first was located at 6p21.31 (referred to as

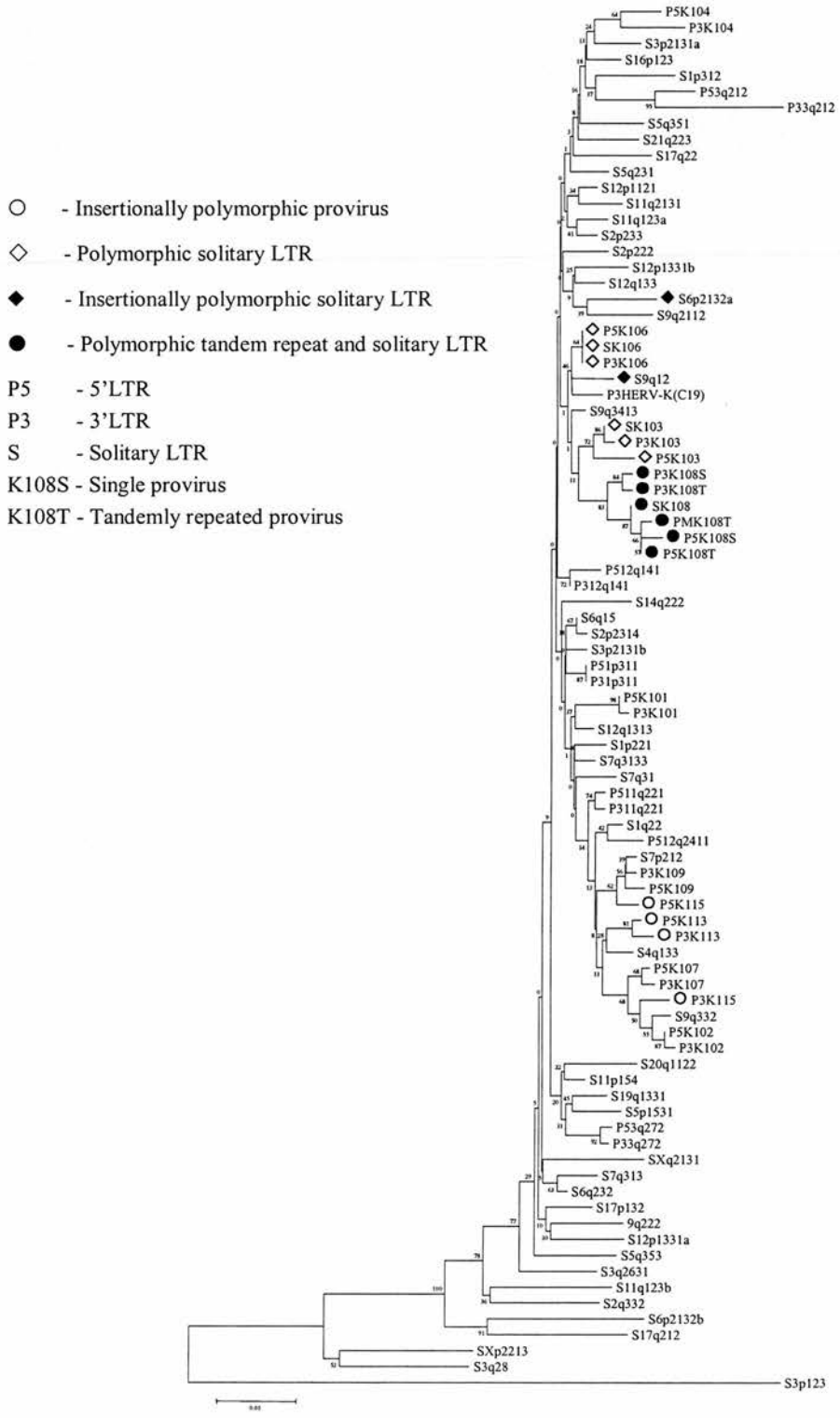
6p21.31a within this study) with accessions Z80898, BX248406, BX927168 and AL662789 containing the solitary LTR and accessions U92032, AL935026 and AL731683, the pre-integration site. All accessions were consistent with the human genomic region 6p21.31, with accessions containing the solitary LTR possessing the sequence ACTTC on either side of the LTR and those analogous to the pre-integration site sequence, one copy of the target site sequence. Further study of the literature regarding the HERV-K(HML-2) LTR at 6p21.31 indicated that the solitary LTR was human specific as a result of the duplication of the MHC (Horton et al., 1998) and not a direct product of the retotransposition of a HERV-K(HML-2) provirus.

The final HERV-K(HML-2) loci to be identified as being polymorphic contained a human specific solitary LTR which was located within the centromeric region of the long arm of chromosome 9 (9q12). Accessions AL39220 and AC013633 consisted of the solitary LTR with the direct repeats CACTG, and accession AL773545 appeared to include the pre-integration sequence which possessed one copy of the target site sequence. It was unclear whether this solitary LTR was insertionally polymorphic as the long arm centromeric region of chromosome 9 appeared to be highly repetitive and the human genome assembly was incomplete for that region.

To investigate the mechanisms that generated allelic variants of the loci; HERV-K103, HERV-K106 and HERV-K108, an alignment of; 49 complete HERV-K(HML-2) human specific solitary LTRs, 30 LTRs belonging to 15 human specific HERV-K(HML-2) proviruses, 3 LTRs contained within HERV-K(HML-2) near complete human specific proviruses, and 5 LTRs belonging to the novel allelic

variants, was constructed using the SIMMONIC sequence alignment package. A neighbour-joining tree was then constructed in MEGA, version 2.1 using the Kimura-2-parameter distance estimate with alignment gaps being handled as a complete deletion and the plylogeny tested with 500 bootstrap replications. The clustering of the novel variant LTRs with the LTRs belonging to their respective progenitor proviruses, indicated that the allelic variants were generated by the process of intra-element recombination (Figure 4.1).

Figure 4.1 Phylogeny of all HERV-K(HML-2) Human Specific LTR sequences and Allelic Variants. The neighbour-joining tree is based on the Kimura-2-parameter distance estimate. Bootstrap values out of 500 re-samplings are shown at the internodes. Individual LTRs are named according to the chromosomal location of the corresponding accession clone or bibliographic name of the sequence.



4.2.2 Collection of Geographically Dispersed DNA Samples

To assess whether the HERV-K(HML-2) loci which were biallelic within the human genome databases could serve as human population genetic markers, DNA samples were collected from 109 unrelated individuals from the geographical regions of Africa, Asia, Europe and Papua-New-Guinea. In addition to the donation of a buccal swab, 75 individuals from Africa, Asia and Europe were asked to provide information regarding their population affiliation, gender and birth place of both parents and themselves. The assembly of these anthropological variables is recommended when performing analysis of human genetic variation as it could augment critical evaluation of single populations (Chakravarti, 1999).

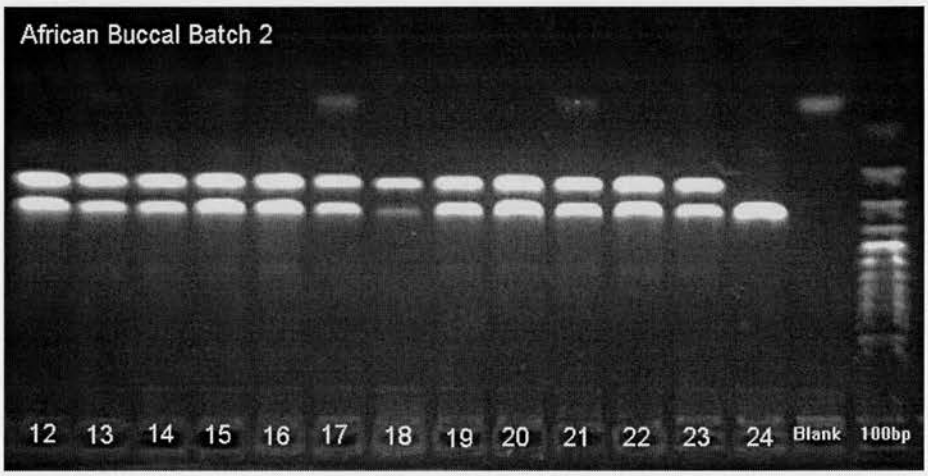
In order to collect non-invasive DNA samples, kits were designed to collect cheek cells in the form of a buccal swab. Prior to the collection of samples, the kits were tested after a three week period to ascertain if following the collection of cheek cells, the swabs required the addition of EDTA to optimise DNA survivability. Amplification for both the first mitochondrial Hypervariable region (HV1) and single copy Amelogenin gene indicated that the addition of EDTA was required for the preservation of single copy DNA (Figure 4.2) (all primers and combinations are listed in Section 2.3.2, Table 2.1).

Following the collection of the anthropologically variable buccal swabs, DNA was extracted from 75 buccal swabs and 34 serum samples (Section 2.2.2), PCR amplification was then performed for the Amelogenin tooth enamel gene which is present on both the X and Y chromosome. The implementation of this amplification assay allowed the quality of DNA template to be assessed and the

Figure 4.2 Survivability of Buccal Swab DNA after a three week period. Wells entitled 'Dry' contain DNA samples that did not have any EDTA added following the collection of cheek cells. Lanes entitled 'Wet' contain swabs that had EDTA added following the collection of the cheek cells. (EBI) - Extraction blank. (EDTA) - EDTA that was added to a buccal swab following collection of cheek cells. (+ ve) - Positive DNA control. (PBl) – Negative DNA control.

gender of the sample to be compared to the recorded information to confirm that samples were not contaminated or mixed up during collection (Figure 4.3). The gender of all buccal samples corresponded to their documented sexual category and did not appear to be contaminated.

Figure 4.3 Amplification for the Amelogenin gene in African swab samples 12 to 24. The Lane entitled 'Blank' contains a negative control. The X chromosome variant produces a fragment of 330 bp and the Y chromosome copy a 218 bp fragment. These representative results show that samples 12 to 23 are from males and sample 24 a female.



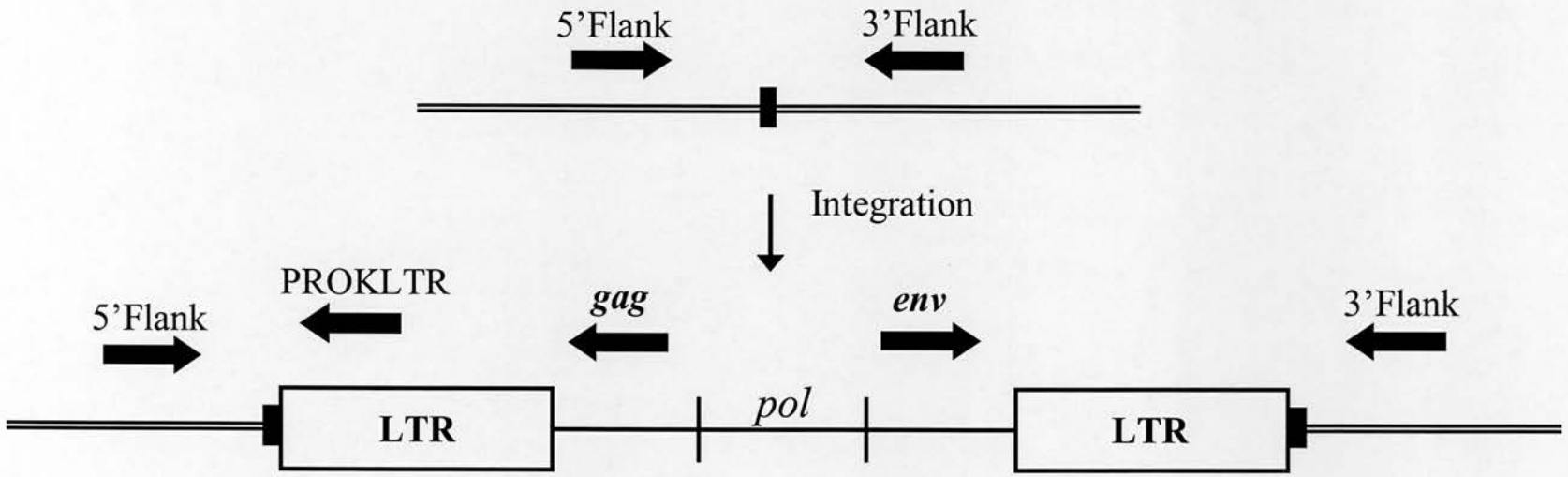
4.2.3 Global Analysis of HERV-K(HML-2) Insertional Proviral Variants

Analysis of the human genetic variation associated with the HERV-K(HML-2) proviral loci that were determined to be biallelic as a result of the presence or absence of insertion (Section 4.2.1), involved the amplification of 109 human DNA samples from four geographical locations. 75 samples were obtained in the form of a buccal swab and 34 in the form of a serum sample. All serum samples were obtained from Papua-New-Guineans. The DNA quality and sexual category of each individual sample was confirmed prior to amplification for the HERV-K(HML-2) biallelic loci (Section 4.2.2).

In order to determine the levels of genetic variation for each HERV-K(HMl-2) locus, a PCR strategy was designed and is visually described in Figure 4.4. DNA sequences adjacent to each proviral insertion were used to design unique flanking region primers. Primers were then screened by standard nucleotide-nucleotide BLAST against the non-redundant and high-thoroughput sequence databases, to ensure that DNA sequences were unique. Universal primers for the LTR, *gag* and *env* regions were designed according to a consensus sequence, which was obtained by aligning all of the HERV-K(HML-2) sequences examined in Section 3.2.1 (for all primers and combinations refer to Section 2.3.2, Table 2.3).

The first amplification reaction involved detection of the pre-integration site sequence and required the unique 5' and 3' flanking region primers (Figure 4.4). The absence of PCR product indicated that either the region had undergone large-scale deletion or that a proviral sequence was present in the homozygous state. For both the HERV-K113 and HERV-K115 provirus, all individuals were positive indicating

Figure 4.4 Location of Primers for the Detection of Biallelic Insertional Proviruses



that they all possessed at least one copy of this allele (Figures 4.5 and 4.6). The second stage of screening was concerned with the amplification for the presence of the HERV-K(HML-2) element at a chosen loci and involved the unique 5' flanking region primer and universal HERV-K LTR antisense primer (Figure 4.4). A positive PCR product indicated the presence of the site specific HERV-K(HML-2) insertion and a negative result implied that the insertion was not present in an individual. In total, 28 individuals were positive for the HERV-K113 insertion and 16 were positive for the insertion of the HERV-K115 sequence (Figures 4.5 and 4.6). In order to confirm that an integrated HERV-K(HML-2) sequence was a complete provirus and not a solitary LTR, amplification was then performed using the unique 5' flanking region primer and universal *gag* antisense primer. As all individuals who were positive for a HERV-K(HML-2) sequence were also PCR positive for this primer combination, they were determined to possess a HERV-K(HML-2) provirus.

To facilitate the analysis of the genetic variation of each of the insertionally polymorphic HERV-K(HML-2) proviruses, the PCR screening results for individuals were complied into their respective geographical location (Tables 4.2 and 4.3). Broad analysis of these results indicates that the African and Papua-New-Guinean populations have a higher frequency of the HERV-K113 and HERV-K115 proviral alleles than the European and Asian populations. In addition, all individuals that possessed a proviral allele were a heterozygote; no individuals were detected to be homozygous for either provirus. These screening results show that the HERV-K113 proviral allele has a worldwide frequency of 0.1386 which indicates that 27 out of 100 people possess this proviral insertion. Likewise, the HERV-K115 proviral allele

Figure 4.5 Amplification for the Pre-Integration Site and HERV-K113 sequence Insertion in African Buccal Swabs 12 to 23. The lane entitled 'Bl' cotanins a negative control. These representative results indicate that samples 16, 18, 20, 22, 23 and 24 are heterozygous and the remaining samples are homozygous for the pre-integration site allele.

5' and 3' Flanking region primers
Pre-integration site

5' Flank and antisense LTR primers
HERV-K(HML-2) insertion

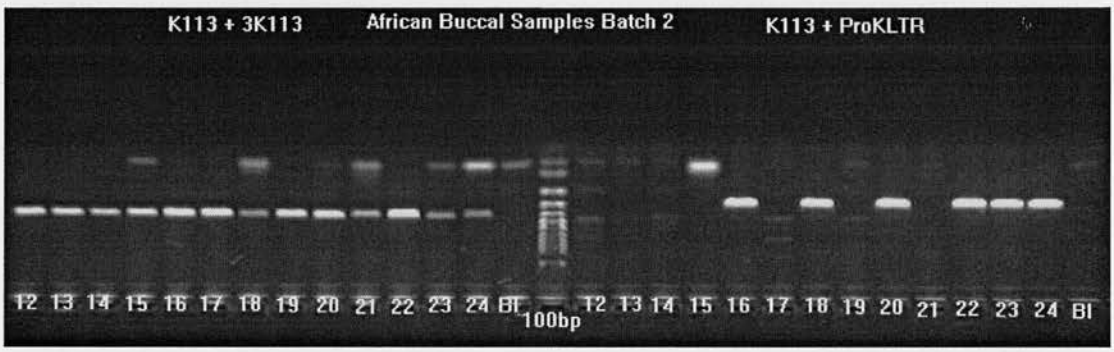


Figure 4.6 Amplification for the Pre-Integration Site and HERV-K115 sequence Insertion in African Buccal Swabs 12 to 23. The lane entitled 'Bl' cotanins a negative control. These representative results indicate that samples 14, 16, 18, 19, 20 and 22 are heterozygous and the remaining samples are homozygous for the pre-integration site allele.

5' and 3' Flanking region primers
Pre-integration site

5' Flank and antisense LTR primers
HERV-K(HML-2) insertion

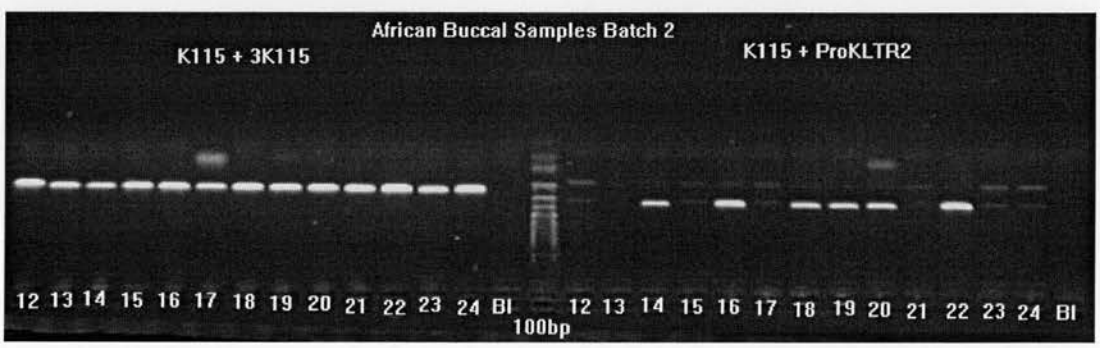


Table 4.2 Human Genomic Diversity of the HERV-K113 Provirus.

(+) Provirus present, (-) Pre-integration site present

Population	Sample(n)	Frequency(+)	(+/+)	(+/-)	(-/-)
Xhosa	2	0.25	0	1	1
Zulu	16	0.281	0	9	7
Tswana	1	0.0	0	0	1
Sierra – Leone	1	0.0	0	0	1
Durban Indian	4	0.0	0	0	4
Swazi	1	0.0	0	0	1
Total Africa	**25**	**0.2**	**0**	**10**	**15**
Chinese	16	0.0937	0	3	13
Korean	1	0.0	0	0	1
Japanese	2	0.25	0	1	1
Javanese	4	0.0	0	0	4
Maduranese	1	0.0	0	0	1
Batak	1	0.0	0	0	1
Achenese	1	0.05	0	1	0
Pakistani	2	0.25	0	1	1
Total Asia	**28**	**0.1071**	**0**	**6**	**22**
Briton	5	0.0	0	0	5
German	3	0.0	0	0	3
French	1	0.0	0	0	1
Swede	1	0.0	0	0	1
Norwegian	1	0.0	0	0	1
Hungarian	1	0.0	0	0	1
Greek	2	0.0	0	0	2
Spanish	2	0.0	0	0	2
Swiss	1	0.0	0	0	1
European Jew	1	0.0	0	0	1
Iraqi Jew	1	0.0	0	0	1
Russian	3	0.0	0	0	3
Total Europe	**22**	**0.0**	**0**	**0**	**22**
Papua-New-Guinea	**26**	**0.2307**	**0**	**12**	**14**
World	**101**	**0.1386**	**0**	**28**	**73**

Table 4.3 Human Genomic Diversity of the HERV-K115 Provirus.

(+) Provirus present, (-) Pre-integration site present

Population	Sample(n)	Frequency(+)	(+/+)	(+/-)	(-/-)
Xhosa	2	0.25	0	1	1
Zulu	16	0.22	0	7	9
Tswana	1	0.5	0	1	0
Sierra – Leone	1	0.0	0	0	1
Durban Indian	4	0.0	0	0	4
Swazi	1	0.5	0	1	0
Total Africa	**25**	**0.2**	**0**	**10**	**15**
Chinese	16	0.0	0	0	16
Korean	1	0.0	0	0	1
Japanese	2	0.0	0	0	2
Javanese	4	0.0	0	0	4
Maduranese	1	0.0	0	0	1
Batak	1	0.0	0	0	1
Achenese	1	0.0	0	0	1
Pakistani	2	0.0	0	0	2
Total Asia	**28**	**0.0**	**0**	**0**	**28**
Briton	5	0.0	0	0	5
German	3	0.0	0	0	3
French	1	0.0	0	0	1
Swede	1	0.0	0	0	1
Norwegian	1	0.0	0	0	1
Hungarian	1	0.0	0	0	1
Greek	2	0.0	0	0	2
Spanish	2	0.0	0	0	2
Swiss	1	0.0	0	0	1
European Jew	1	0.0	0	0	1
Iraqi Jew	1	0.0	0	0	1
Russian	3	0.0	0	0	3
Total Europe	**22**	**0.0**	**0**	**0**	**22**
Papua-New-Guinea	**34**	**0.0882**	**0**	**6**	**28**
World	**109**	**0.0734**	**0**	**16**	**93**

has a worldwide frequency of 0.1467 indicating that 7 people in every 50 are expected to possess the provirus.

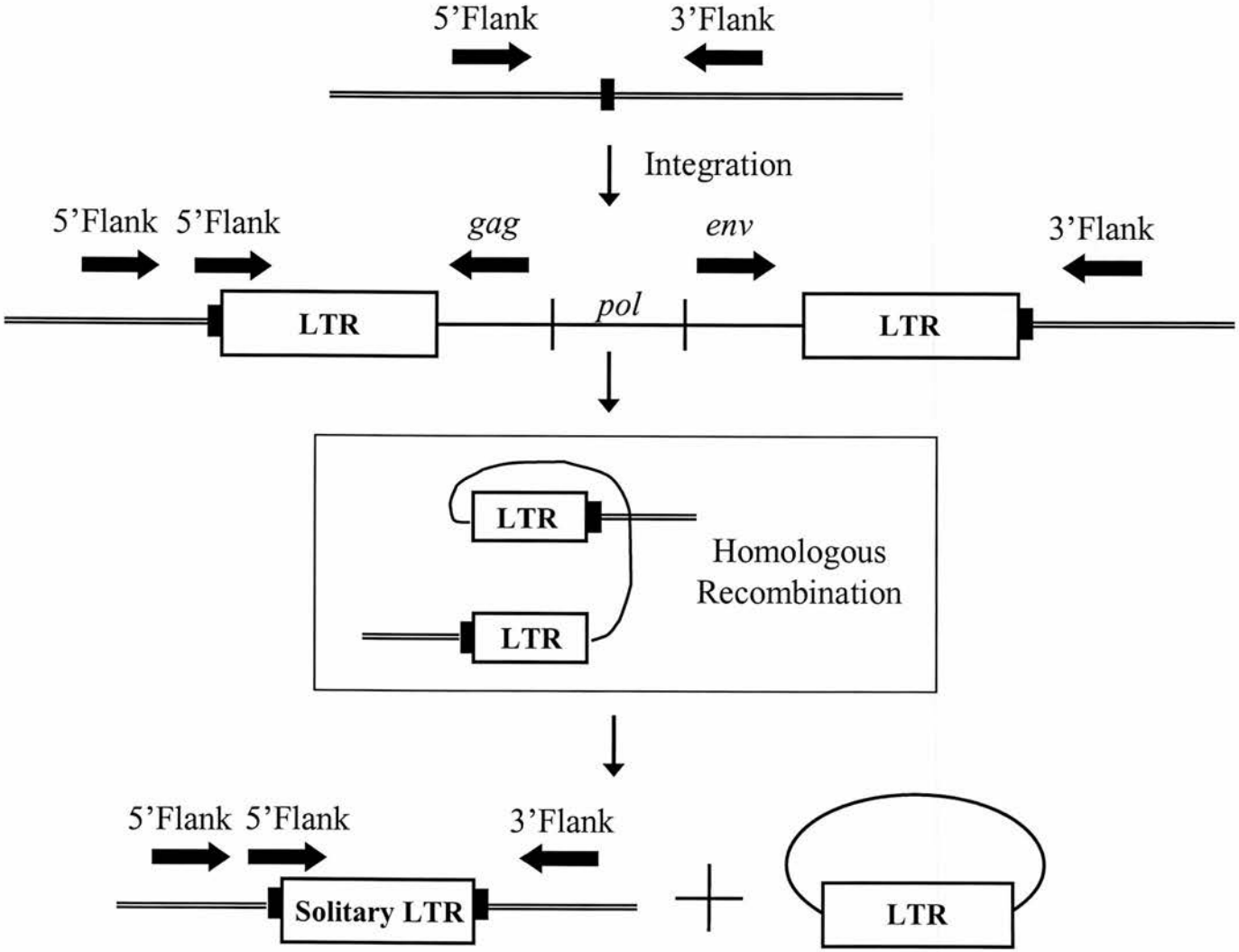
4.2.4 Global Analysis of HERV-K(HML-2) Solitary LTR Variants

Investigation of the human genetic variation associated with the HERV-K(HML-2) proviral loci that were determined to be biallelic as a result of the presence of a complete provirus or solitary LTR (Section 4.2.1), involved the amplification of 109 human DNA samples from four geographical locations. Thirty four serum samples were obtained from Papua-New-Guineans, 25 from Africans, 28 from Asians and 22 from Europeans. The quality of DNA and sexual category of each individual was confirmed prior to amplification for the HERV-K(HML-2) biallelic loci (Section 4.2.2).

To determine the levels of genetic variation for each HERV-K(HML-2) locus, a PCR strategy was designed and is visually described in Figure 4.7. DNA sequences adjacent to each proviral insertion were used to design unique flanking region primers. Primers were then screened by standard nucleotide-nucleotide BLAST against the non-redundant and high-thoroughput sequence databases, to ensure that DNA sequences were unique. Universal primers for the LTR, *gag* and *env* regions were designed according to a consensus sequence, which was obtained by aligning all of the HERV-K(HML-2) sequences examined in Section 3.2.1 (all primers and combinations are listed in Section 2.3.2, Table 2.3).

The first amplification reaction involved detection of the HERV-K proviral sequence and required the unique 5' flanking region primer and universal *gag* antisense primer (Figure 4.7). A negative PCR result indicated either the individual did not possess the HERV-K proviral sequence, they were homozygous for another allele or that the DNA sample was degraded. As the single round amplification for

Figure 4.7 Location of Primers for the Detection of Biallelic Solitary LTRs



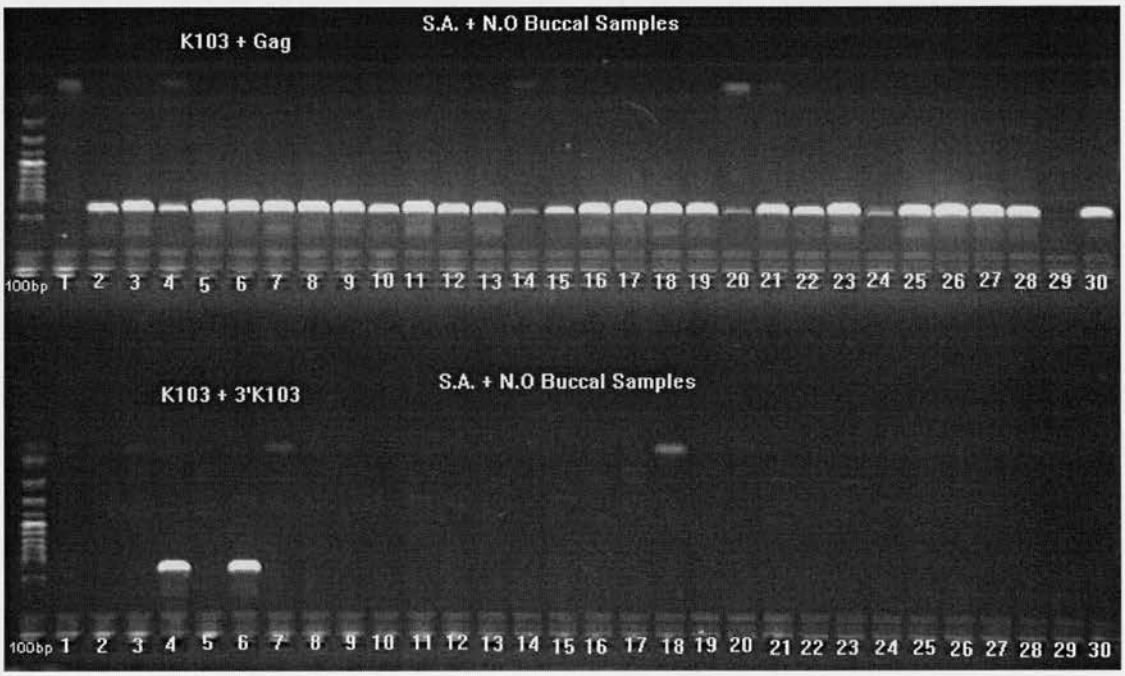
the HERV-K106 provirus proved difficult to optimise, hemi-nested PCR reactions were performed with a second 5' flanking region primer which covered the start of the provirus. For the HERV-K103 provirus, 90 individuals were PCR positive and 101 individuals were positive for the HERV-K106 provirus (Figures 4.8 and 4.9a). These individuals were therefore determined to possess at least one copy of the proviral allele.

The second stage of screening was concerned with the amplification for the presence of the HERV-K(HML-2) solitary LTRs and involved using the unique 5' and 3' flanking region primers. A negative PCR result indicated either that an individual possessed a pre-integration site allele, they were homozygous for the proviral allele or the sample was degraded. As with the amplification for the HERV-K106 provirus, hemi-nested reactions involving the 5' flanking region primer, which covered the beginning of the provirus, were performed for the solitary LTR as single round amplification proved difficult to optimise (Figure 4.7). In total, only two individuals belonging to the geographical region of Africa were PCR positive for the HERV-K103 solitary LTR (Figure 4.8). For the HERV-K106 solitary LTR, 15 individuals were positive for the allele (Figure 4.9b), including one Papua-New-Guinean individual who was PCR negative for the HERV-K106 provirus, indicative of either the individual being homozygous for the solitary LTR or that they also possessed an allele for the pre-integration site.

As 10 out of 109 DNA samples did not produce an amplicon for either the HERV-K 103 provirus or solitary LTR and 8 out of 109 samples were negative for the HERV-K106 provirus or solitary LTR, a third stage of screening was performed. This involved a single round of amplification using the outer unique 5' and 3'

Figure 4.8 Amplification for the HERV-K103 Provirus and Solitary LTR in African Buccal Swabs 1 to 24 (Lanes 1 to 24). Lane 29 contains an extraction blank and lane 30 a positive control. These representative results indicate that samples 4 and 6 are heterozygous and the remaining samples are homozygous.

(Below) Use of the 5' flank and antisense *gag* primers to amplify the proviral allele



(Above) Use of the 5' and 3' flanking region primers to amplify a solitary LTR

Figure 4.9a Amplification for the HERV-K106 Provirus in Papua-New-Guinean Serum Samples 17 to 34 (lanes 1 to 17). Lane 18 contains a negative control and lane 19 a positive control. These representative results indicate that all samples, with the exclusion of sample 28 (lane 11), possess at least one copy of the HERV-K106 provirus.

Use of the 5' flank and antisense *gag* primers to amplify the proviral allele

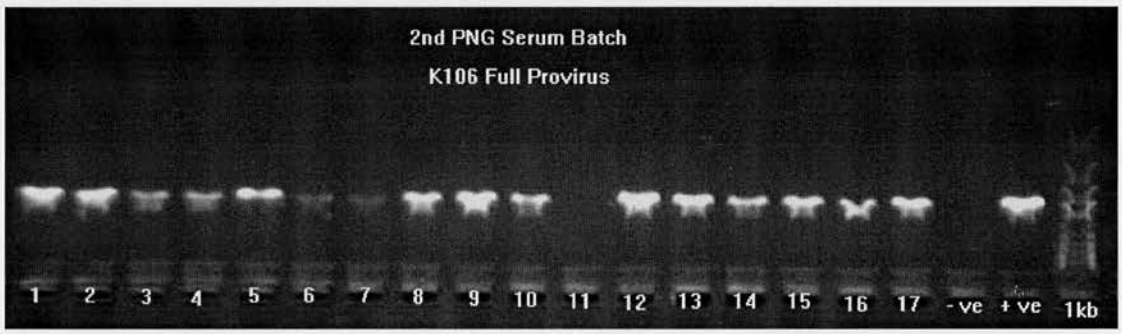
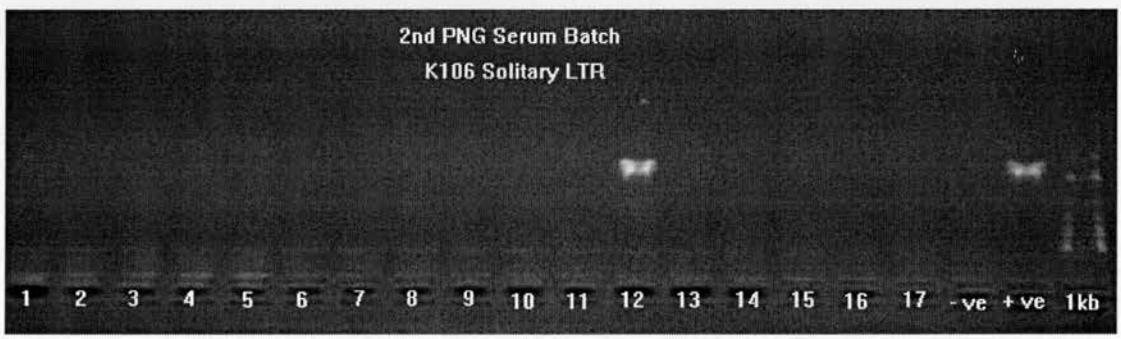


Figure 4.9b Amplification for the HERV-K106 Solitary LTR in Papua-New-Guinean Serum Samples 17 to 34 (Lanes 1 to 17). Lane 18 contains a negative control and lane 19 a positive control. These representative results indicate that sample 29 (lane 12) is positive for the solitary LTR and so a heterozygote. All remaining individuals, with the exception of sample 28 (lane 11), are homozygous for the HERV-K106 provirus.

Use of the 5' and 3' flanking region primers to amplify a solitary LTR



flanking region primers and a shorter extension time than the second stage of screening (Figure 4.7). If a positive PCR product was generated, it would indicate that a third allele representing the pre-integration site was present within contemporary human populations. Alternatively, if a sample was negative for the pre-integration site and did not produce an amplicon for the two alleles present within the human genome databases, it could be ascertained that the DNA sample was degraded. All samples which were PCR negative for the proviral and solitary LTR alleles were also PCR negative for the pre-integration site, this implied that the DNA quality within these individual samples was inadequate for analysis.

To facilitate the analysis of the genetic variation of each of the structurally variable HERV-K(HML-2) proviral loci, the individual screening results were complied into their respective geographical location (Tables 4.4 and 4.5). Broad analysis of these results indicates that the African population has the highest frequency of both the HERV-K103 and HERV-K106 solitary LTR alleles. With the exception of one Papua-New-Guinean individual, all individuals who possessed a solitary LTR allele were heterozygous. These screening results show that the HERV-K103 solitary LTR has a worldwide frequency of 0.011 which indicates that 1 in 100 people will possess the solitary LTR (Table 4.4). Likewise, the HERV-K106 solitary LTR allele has a worldwide frequency of 0.07425 indicating that 7 people in every 100 will possess the solitary LTR (Table 4.5).

Table 4.4 Human Genomic Diversity of the HERV-K103 Solitary LTR.

(+) solitary LTR present, (-) complete provirus present

Population	Sample(n)	Frequency(+)	(+/+)	(+/-)	(-/-)
Xhosa	2	0.0	0	0	2
Zulu	16	0.06	0	2	14
Tswana	1	0.0	0	0	1
Sierra – Leone	1	0.0	0	0	1
Durban Indian	4	0.0	0	0	4
Swazi	1	0.0	0	0	1
Total Africa	**25**	**0.04**	**0**	**2**	**23**
Chinese	16	0.0	0	0	16
Korean	1	0.0	0	0	1
Japanese	2	0.0	0	0	2
Javanese	4	0.0	0	0	4
Maduranese	1	0.0	0	0	1
Batak	1	0.0	0	0	1
Achenese	1	0.0	0	0	1
Pakistani	2	0.0	0	0	2
Total Asia	**28**	**0.0**	**0**	**0**	**28**
Briton	5	0.0	0	0	5
German	3	0.0	0	0	3
French	1	0.0	0	0	1
Swede	1	0.0	0	0	1
Norwegian	1	0.0	0	0	1
Hungarian	1	0.0	0	0	1
Greek	2	0.0	0	0	2
Spanish	2	0.0	0	0	2
Swiss	1	0.0	0	0	1
European Jew	1	0.0	0	0	1
Iraqi Jew	1	0.0	0	0	1
Russian	3	0.0	0	0	3
Total Europe	**22**	**0.0**	**0**	**0**	**22**
Total Papua-New-Guinea	**15**	**0.0**	**0**	**0**	**15**
World	**90**	**0.011**	**0**	**2**	**88**

Table 4.5 Human Genomic Diversity of the HERV-K106 Solitary LTR.

(+) solitary LTR present, (-) complete provirus present

Population	Sample(n)	Frequency(+)	(+/+)	(+/-)	(-/-)
Xhosa	2	0.0	0	0	2
Zulu	16	0.15	0	5	11
Tswana	1	0.0	0	0	1
Sierra – Leone	1	0.0	0	0	1
Durban Indian	4	0.0	0	0	4
Swazi	1	0.0	0	0	1
Total Africa	**25**	**0.1**	**0**	**5**	**20**
Chinese	16	0.0625	0	2	14
Korean	1	0.0	0	0	1
Japanese	2	0.0	0	0	2
Javanese	3	0.333	0	2	1
Maduranese	1	0.0	0	0	1
Batak	1	0.0	0	0	1
Achenese	1	0.0	0	0	1
Pakistani	2	0.0	0	0	2
Total Asia	**27**	**0.074**	**0**	**4**	**23**
Briton	5	0.1	0	1	4
German	3	0.0	0	0	3
French	1	0.0	0	0	1
Swede	1	0.5	0	1	0
Norwegian	1	0.0	0	0	1
Hungarian	1	0.0	0	0	1
Greek	2	0.0	0	0	2
Spanish	2	0.0	0	0	2
Swiss	1	0.0	0	0	1
European Jew	1	0.5	0	1	0
Iraqi Jew	1	0.0	0	0	1
Russian	2	0.0	0	0	2
Total Europe	**21**	**0.0714**	**0**	**3**	**18**
Total Papua-New-Guinea	**28**	**0.0714**	**1**	**2**	**25**
World	**101**	**0.07425**	**1**	**14**	**86**

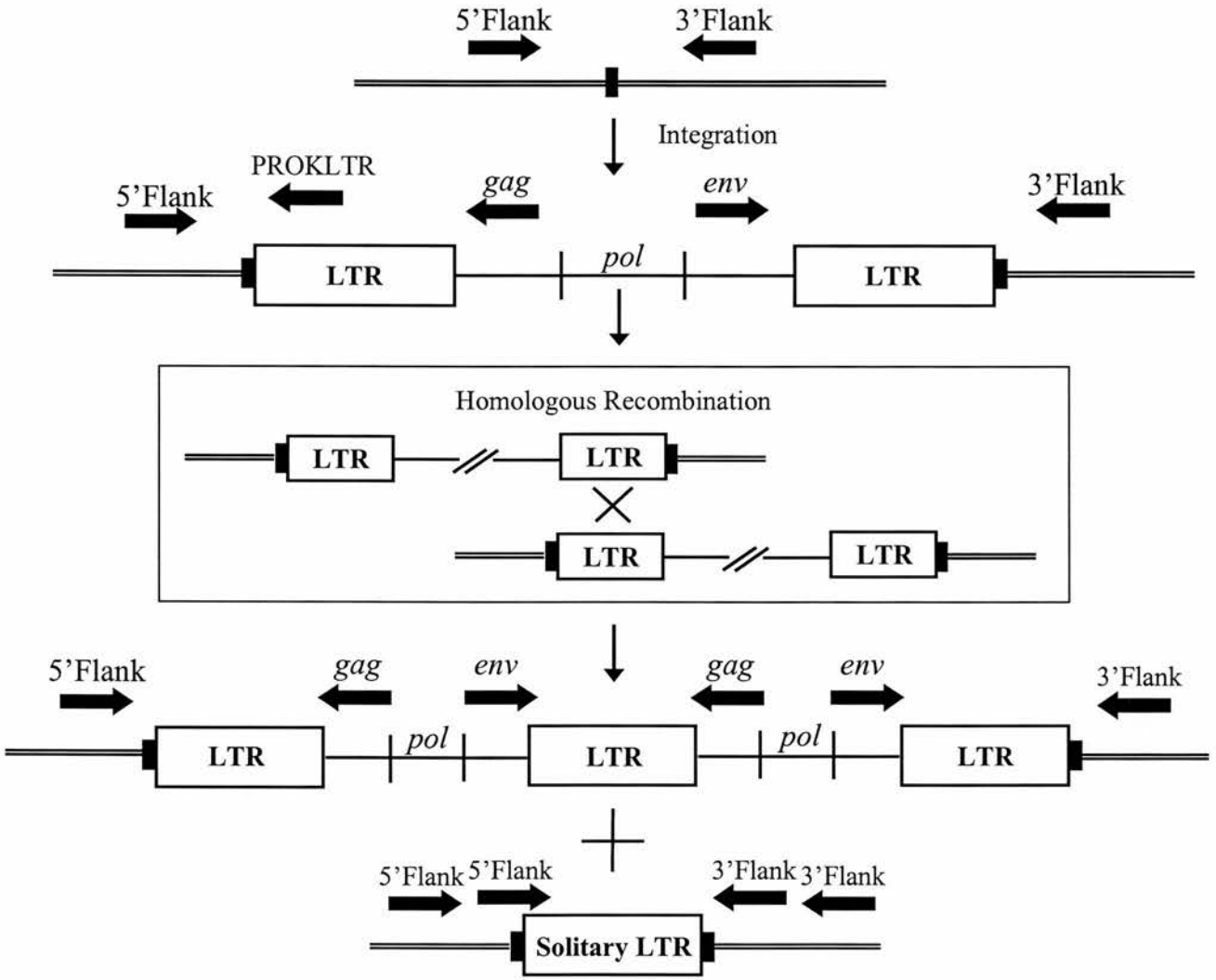
4.2.5 Global Analysis of the HERV-K108 Tandem Repeat and Solitary LTR

Determination of the human genetic variation associated with the HERV-K(HML-2) proviral loci HERV-K108, which was established to be biallelic as a result of the presence or absence of a tandem repeat (Section 4.2.1), involved the amplification of 109 human DNA samples from four geographical locations. 75 samples were obtained in the form of a buccal swab and 34 in the form of a serum sample. All serum samples belonged to Papua-New-Guineans and the remaining 75 buccal samples from Africans, Europeans and Asians. The DNA quality and sexual category of each individual sample was confirmed prior to amplification for the HERV-K(HML-2) biallelic loci (Section 4.2.2).

A PCR strategy was designed in order to determine the levels of global genetic variation of the HERV-K108 locus, which is visually described in Figure 4.10. DNA sequences adjacent to each proviral insertion were used to design unique flanking region primers. Primers were then screened by standard nucleotide-nucleotide BLAST against the non-redundant and high-thoroughput sequence databases, to ensure that DNA sequences were unique. Universal primers for the LTR, *gag* and *env* regions were designed according to a consensus sequence, which was obtained by aligning all of the HERV-K(HML-2) sequences examined in Section 3.2.1 (all primers and combinations are listed in Section 2.3.2, Table 2.3)

The first stage of screening involved the detection for the presence of a proviral sequence at the locus, using the unique 5' flanking region primer and universal antisense *gag* primer (Figure 4.10). Ninety-one out of 109 individuals were

Figure 4.10 Location of Primers for the Detection of the HERV-K108 Tandemly Duplicated Provirus and Solitary LTR



PCR positive indicating that they possessed at least one copy of either the HERV-K108 single provirus or tandemly repeated provirus (Figure 4.11).

The second stage of testing was concerned with the detection of the tandemly repeated provirus and involved the universal antisense *gag* primer and sense *env* primer. Computational screening within the human genome databases for the potential combinations of these universal primers indicated that the predicted amplicon was unique to the HERV-K108 locus on chromosome 7. In total 46 out of 109 individuals screened possessed at least one copy of the tandem repeat (Figure 4.12).

Homologous recombination between sister chromatids at a region containing a HERV provirus (intra-element recombination) is expected to generate both a tandem repeat and solitary LTR (Figure 4.10). As a solitary LTR was not detected at this locus within the human genome databases (Section 4.2.1), individuals were also screened for the possession of a solitary LTR to confirm if the tandem repeat was generated by an intra-element recombination event. Nested reactions involving two 3' flanking region primers and two 5' flanking region primers, one of which covered the beginning of the proviral sequence, were performed as single round PCR proved difficult to optimise (Figure 4.10). In total, only one African individual was PCR positive for this allele (Figure 4.13). As this individual was previously PCR positive for the HERV-K108 proviral sequence but negative for the tandem repeat, it was determined that they were heterozygous for a single provirus and solitary LTR.

In view of the fact that only 91 out of 109 individuals were PCR positive for the HERV-K108 proviral sequence and only one possessed the solitary LTR, amplification was then performed for the HERV-K108 pre-integration site sequence

Figure 4.11 Amplification for the HERV-K108 Proviral sequence using a 5' flanking region and antisense *gag* primer, in African Buccal Swabs 1 to 11 (Lanes 1 to 11). Lane 12 contained an extraction blank and lane 13 a positive control. These representative results indicate that all individuals possess at least one copy of the HERV-K108 proviral sequence.

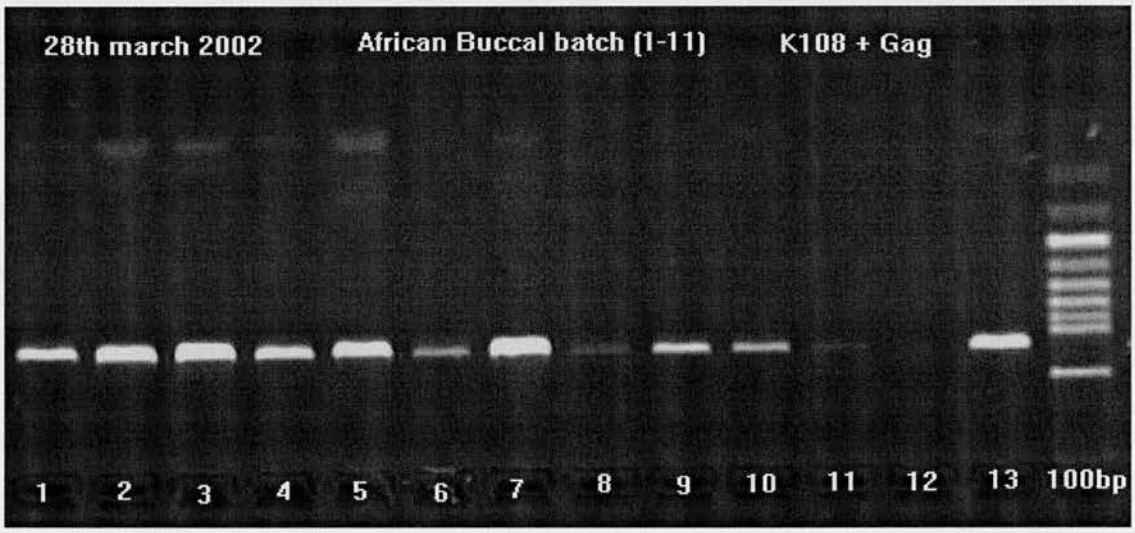


Figure 4.12 Amplification for the HERV-K108 Tandemly Repeated Provirus using the universal *env* and antisense *gag* primers, in African Buccal Swabs 1 to 11 (Lanes 1 to 11). Lane 12 contained an extraction blank and lane 13 a positive control. These representative results indicate that samples 1, 3, 6, 7 and 11 possess at least one copy of the tandem repeat.

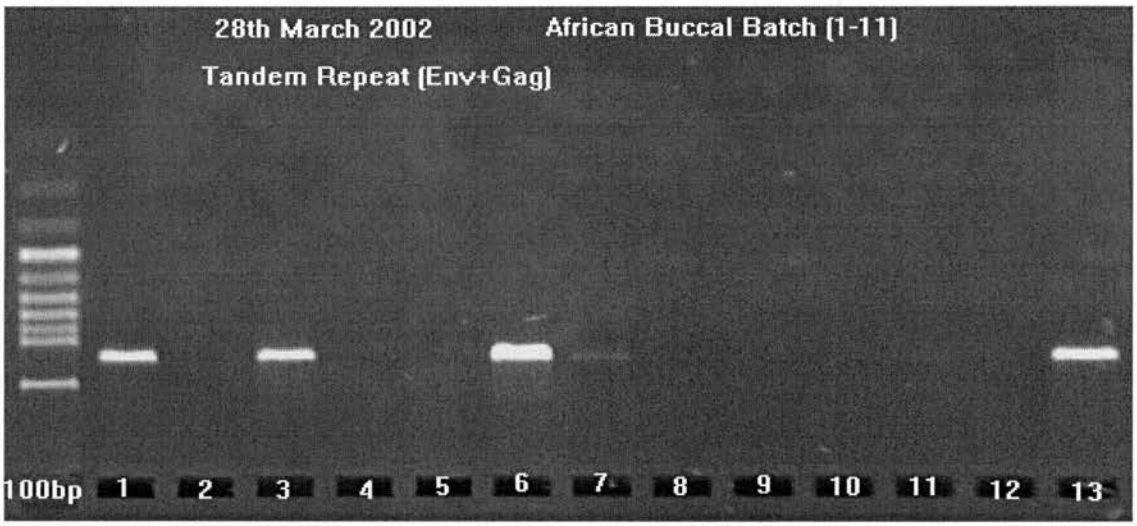
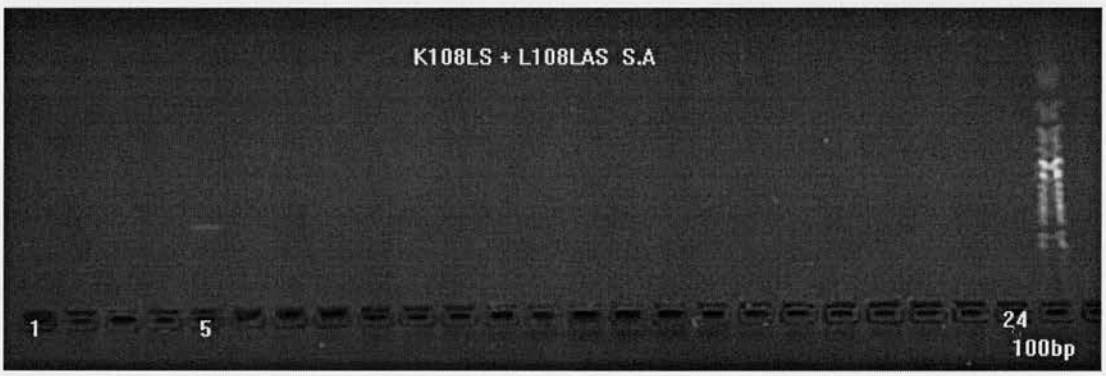


Figure 4.13 Amplification for the HERV-K108 Solitary LTR African Buccal Swabs 1 to 24 (Lanes 1 to 24), by the application of a Fully Nested PCR Assay. The first round utilised unique 5' and 3' flanking region primers. The second round reaction used a 5' flanking region primer which covered part of the HERV-K108 insertion and a unique 3' flanking region primer which was internal to the first 3' primer. These representative results indicate that sample 5 possesses at least one copy of the solitary LTR. As this individual was previously determined to be positive for the HERV-K108 proviral sequence but negative for the tandem repeat, it was determined that they were heterozygous for a single provirus and solitary LTR



to determine if a fourth allele was present at this locus utilising the unique 5' and 3' flanking region primers (Figure 4.10). All individuals were PCR negative indicating that the pre-integration site was not a forth allele within human populations. To ensure that the negative outcome for the HERV-K108 proviral sequence was due to poor quality of template, reactions were also performed for the 3' end of the proviral sequence using the antisense 3' flanking region primer and sense *env* primer. As with the screening for the 5' proviral sequence, all samples were negative.

To assist the statistical analysis of the genetic variation of the HERV-K108 locus, the PCR screening results for individuals were compiled into their respective geographical location (Table 4.6). As amplification reactions that spanned the entire length of the HERV-K108 locus were not performed as products would have been either ~ 10 kb or ~ 19 kb in length. It was impossible to distinguish between individuals who were heterozygous in possessing one copy of the ancestral single proviral allele (A) and a copy of the tandemly repeated provirus (B), from individuals who were homozygous for the tandemly repeated provirus (B). However, in performing conformational amplification for the presence of the HERV-K108 proviral sequence, the number of individuals who were homozygous in possessing the single proviral could be determined (AA). Overall, it appeared that approximately half of the individuals tested possessed a copy of the tandem repeat with the remaining individuals being homozygous for the single provirus (Table 4.6). The expected reciprocal product of intra-element homologous recombination, a solitary LTR, was only present in a single individual, indicative of an allele frequency of 0.02 within the African population and a worldwide frequency of 0.005. This would imply that only 1 in 200 people will possess this allele.

Table 4.6 Human Genomic Diversity of the HERV-K108 locus.

(A) Single Provirus, (B) Tandem Repeat, (C) Solitary LTR)

Population	Sample(n)	(AA)	(AB/BB)	(AC)
Xhosa	2	1	1	0
Zulu	16	5	11	0
Tswana	1	0	0	1
Sierra – Leone	1	0	1	0
Durban Indian	4	2	2	0
Swazi	1	0	1	0
Total Africa	**25**	**8**	**16**	**1**
Chinese	16	12	4	0
Korean	1	0	1	0
Japanese	2	1	1	0
Javanese	4	2	2	0
Maduranese	1	1	0	0
Batak	1	0	1	0
Achenese	1	1	0	0
Pakistani	2	0	2	0
Total Asia	**28**	**17**	**11**	**0**
Briton	5	1	4	0
German	3	0	3	0
French	1	1	0	0
Swede	1	0	1	0
Norwegian	1	0	1	0
Hungarian	1	1	0	0
Greek	2	1	1	0
Spanish	2	1	1	0
Swiss	1	0	1	0
European Jew	1	0	1	0
Iraqi Jew	1	0	1	0
Russian	2	2	0	0
Total Europe	**21**	**7**	**14**	**0**
Total Papua-New-Guinea	**17**	**12**	**5**	**0**
World	**91**	**44**	**46**	**1**

To further examine the aberrant events that generated the solitary LTR, African individual sample 5 was sequenced and the solitary compared to the other HERV-K108 LTRs (Figure 4.14). According to diagnostic nucleotide differences it appears that a potential crossover point could be between 290 bp and 340 bp from the beginning of the LTR. This implies that the beginning of the solitary LTR is 5' LTR-like and then 3' LTR-like. Similarly, the central LTR of the tandemly repeated provirus, appears to be 3'LTR-like and 5'LTR-like. In combination with the grouping of the respective HERV-K108 LTRs in the phylogenetic tree presented as Figure 4.1 in Section 4.2.1, a reciprocal homologous recombination event such as illustrated in Figure 4.10, is likely to have generated the novel HERV-K108 structural allelic variants.

Figure 4.14 Alignment of HERV-K108 LTRs belonging to different alleles.

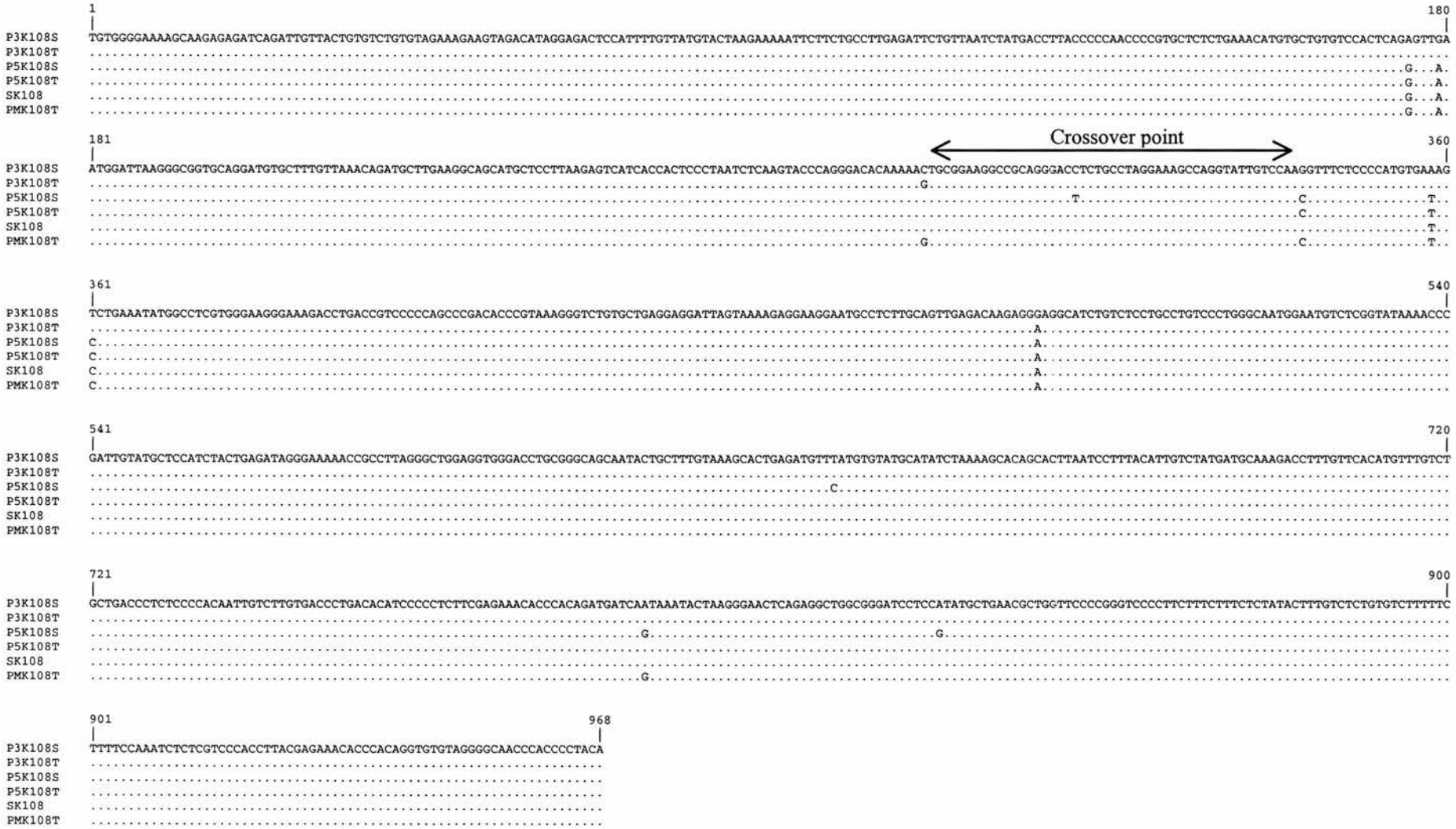


Figure 4.14 Alignment of HERV-K108 LTRs belonging to different alleles. (P5) 5'LTR, (P3) 3'LTR, (K108S) Part of the single copy provirus, (K108T) Part of the tandemly repeated provirus, (SK108) Solitary LTR, (PMK108T) Middle LTR of the tandem repeat. Nucleotide substitutions at each position are indicated with the appropriate nucleotide.

4.2.6 Statistical Analysis of HERV-K(HML-2) Allelic Variants

A total of 109 individuals from four geographically dispersed populations were screened for five polymorphic HERV-K(HML-2) loci. For each locus, excluding the HERV-K108 region, the allele frequency of each variant was calculated within the PopGene statistical package and is presented in Tables 4.7 to 4.10. Frequencies for the four major population groups range from 0.231 for the HERV-K113 provirus in the Papua New Guinean population to zero for the novel variant in the European sample (Table 4.7). The allele frequency of the HERV-K115 provirus is highest in the African sample with a frequency of 0.2 and lowest in the Asian and European population groups, which only contain individuals who are homozygous for the pre-integration site allele (Table 4.8). Allele frequencies for the HERV-K103 solitary LTR range from 0.04 in the African population to zero in all other populations, on a worldwide scale, the expected frequency of the solitary LTR is 0.011 (Table 4.9). The HERV-K106 solitary LTR is present in all of the four populations tested, with allele frequencies ranging from 0.1 in the African sample to 0.048 in the European sample with a subsequent global frequency of 0.069 (Table 4.10). This indicates that the HERV-K106 solitary LTR allele is geographically more widespread than the HERV-K103 solitary LTR. Overall, average allele frequencies for the novel alleles are highest in the African sample and lowest in the European population.

Tests of the Hardy-Weinberg equilibrium, which predicts that in a randomly mating population in the absence of selection and migration, allele frequencies remain consistent from one generation to the next, were calculated in the PopGene

Table 4.7 Allele Frequencies, Heterozygosity, and F_{ST} of HERV-K113. (A) Pre-integration site allele, (B) Complete provirus allele.

Population	Sample Size	Allele	Allele Frequency	Observed Heterozygosity	Expected Heterozygosity	F_{ST}
Africa	25	A	0.8	0.4	0.32	
		B	0.2			
Asia	28	A	0.892	0.214	0.191	
		B	0.108			
Europe	22	A	1.0	0.0	0.0	
		B	0.0			
PNG	26	A	0.769	0.461	0.355	
		B	0.231			
Global	101	A	0.861	0.277	0.238	0.0696
		B	0.139			

Table 4.8 Allele Frequencies, Heterozygosity, and F_{ST} of HERV-K115. (A) Pre-integration site allele, (B) Complete provirus allele.

Population	Sample Size	Allele	Allele Frequency	Observed Heterozygosity	Expected Heterozygosity	F_{ST}
Africa	25	A	0.8	0.4	0.32	
		B	0.2			
Asia	28	A	1.0	0.0	0.0	
		B	0.0			
Europe	22	A	1.0	0.0	0.0	
		B	0.0			
PNG	34	A	0.897	0.205	0.184	
		B	0.103			
Global	109	A	0.922	0.156	0.143	0.0988
		B	0.078			

Table 4.9 Allele Frequencies, Heterozygosity, and F_{ST} of HERV-K103. (A) Complete provirus allele, (B) Solitary LTR allele.

Population	Sample Size	Allele	Allele Frequency	Observed Heterozygosity	Expected Heterozygosity	F_{ST}
Africa	25	A	0.96	0.08	0.076	
		B	0.04			
Asia	28	A	1.0	0.0	0.0	
		B	0.0			
Europe	22	A	1.0	0.0	0.0	
		B	0.0			
PNG	15	A	1.0	0.0	0.0	
		B	0.0			
Global	90	A	0.989	0.021	0.021	0.0303
		B	0.011			

Table 4.10 Allele Frequencies, Heterozygosity, and F_{ST} of HERV-K106. (A) Complete provirus allele, (B) Solitary LTR allele.

Population	Sample Size	Allele	Allele Frequency	Observed Heterozygosity	Expected Heterozygosity	F_{ST}
Africa	25	A	0.9	0.2	0.18	
		B	0.1			
Asia	27	A	0.944	0.111	0.104	
		B	0.056			
Europe	21	A	0.952	0.095	0.090	
		B	0.048			
PNG	28	A	0.931	0.069	0.128	
		B	0.069			
Global	101	A	0.931	0.117	0.127	0.0063
		B	0.069			

statistical package using the algorithm by Levene, (1949) and are presented in Tables 4.11 to 4.14. Tests were performed for each locus in the populations of Africa, Asia, Europe and Papua-New Guinea, with tests of fit also being calculated on a global scale. A total of 19 tests out of 20 were significant for the χ^2 (Chi-squared) goodness of fit test at the 5 %, 1 % and 0.1 % level percentiles, which provides strong statistical support for all populations fitting within the Hardy-Weinberg Law. The Papua-New-Guinean population departed from the Hardy-Weinberg equilibrium at the HERV-K106 locus at the 5 % and 1 % levels, however the population remains within the Law as the χ^2 test remains significant at the 0.1 % level with a score of 8.482 (Table 4.14).

As a small population sample size may not adequately represent the frequency of all genotypes within a population, likelihood ratio tests (G^2) were also calculated using the PopGene statistical package, which applies the algorithm by Levene, (1949), to confirm if all populations did fit within the Hardy-Weinberg Law (Tables 4.11 to 4.14). The Null hypothesis was proven to be correct in 20 out of 20 tests, including the HERV-K106 locus for the Papua-New-Guinean population, at the 5 %, 1 % and 0.1 % level percentiles.

Observed and expected heterozygosity of all loci and populations was calculated using the PopGene statistical package, which employs the methods of Levene, (1949) and Nei, (1973), with results presented in Tables 4.7 to 4.10. The heterozygosity for the HERV-K113 provirus varied from 0.355 for the Papua-New Guineans to 0.32 in the African population and was 0.0 in both the Asian and European samples as they were both monomorphic for the pre-integration site allele (Table 4.7). For the HERV-K115 provirus, heterozygosity ranged from 0.32 for the

Table 4.11 Hardy-Weinberg equilibrium tests for the HERV-K113 locus. (A) Pre-integration site allele, (B) Complete provirus allele.

Population	Genotypes	Observed (O)	Expected (E)	$(O-E)^2 / E$	χ^2 (df = 1)	Probability	G^2 (df = 1)	Probability
Africa	AA	15	15.918	0.053	1.384	0.239	2.276	0.131
	AB	10	8.163	0.413				
	BB	0	0.918	0.918				
Asia	AA	22	22.272	0.003	0.33	0.565	0.601	0.437
	AB	6	5.454	0.054				
	BB	0	0.272	0.272				
Europe	AA	Monomorphic						
	AB							
	BB							
PNG	AA	14	15.294	0.109	2.115	0.145	3.355	0.066
	AB	12	9.411	0.711				
	BB	0	1.294	1.294				
Global	AA	73	74.880	0.047	2.511	0.113	4.364	0.036
	AB	28	24.238	0.583				
	BB	0	1.880	1.880				

Table 4.12 Hardy-Weinberg equilibrium tests for the HERV-K115 locus. (A) Pre-integration site allele, (B) Complete provirus allele.

Population	Genotypes	Observed (O)	Expected (E)	$(O-E)^2 / E$	χ^2 (df = 1)	Probability	G^2 (df = 1)	Probability
Africa	AA	15	15.918	0.053	1.384	0.239	2.276	0.131
	AB	10	8.163	0.413				
	BB	0	0.918	0.918				
Asia	AA	Monomorphic						
	AB							
	BB							
Europe	AA	Monomorphic						
	AB							
	BB							
PNG	AA	27	27.313	0.003	0.378	0.538	0.690	0.406
	AB	7	6.373	0.061				
	BB	0	0.313	0.313				
Global	AA	92	92.626	0.004	0.730	0.392	1.354	0.244
	AB	17	15.746	0.099				
	BB	0	0.626	0.626				

Table 4.13 Hardy-Weinberg equilibrium tests for the HERV-K103 locus. (A) Complete provirus allele, (B) Solitary LTR allele.

Population	Genotypes	Observed (O)	Expected (E)	$(O-E)^2 / E$	χ^2 (df = 1)	Probability	G^2 (df = 1)	Probability
Africa	AA	23	23.02	0.0	0.021	0.884	0.416	0.838
	AB	2	1.959	0.0				
	BB	0	0.02	0.02				
Asia	AA	Monomorphic						
	AB							
	BB							
Europe	AA	Monomorphic						
	AB							
	BB							
PNG	AA	Monomorphic						
	AB							
	BB							
Global	AA	91	91.005	0.00	0.005	0.941	0.010	0.916
	AB	2	1.989	0.0001				
	BB	0	0.005	0.005				

Table 4.14 Hardy-Weinberg equilibrium tests for the HERV-K106 locus. (A) Complete provirus allele, (B) Solitary LTR allele.

Population	Genotypes	Observed (O)	Expected (E)	$(O-E)^2 / E$	χ^2 (df = 1)	Probability	G^2 (df = 1)	Probability
Africa	AA	20	20.204	0.002	0.242	0.622	0.445	0.504
	AB	5	4.591	0.036				
	BB	0	0.204	0.204				
Asia	AA	24	24.056	0.0001	0.061	0.804	0.117	0.731
	AB	3	2.886	0.004				
	BB	0	0.566	0.056				
Europe	AA	19	19.024	0.00	0.025	0.872	0.05	0.823
	AB	2	1.951	0.001				
	BB	0	0.024	0.024				
PNG	AA	26	25.105	0.031	8.482	0.003	3.767	0.052
	AB	2	3.789	0.845				
	BB	1	0.105	7.605				
Global	AA	89	88.448	0.003	0.775	0.378	0.6	0.438
	AB	12	13.103	0.092				
	BB	1	0.448	0.679				

African population to 0.0 within the monomorphic Asian and European populations (Table 4.8). The measure of heterozygosity of the HERV-K103 locus is exclusively represented by the African population, with a score of 0.076, as the novel solitary LTR allele is only present in two African individuals (Table 4.9). The heterozygosity of the HERV-K106 locus varied from 0.18 in the African population to 0.09 in the European sample (Table 4.10). All observed and expected heterozygosity values are equivalent with the exception of the HERV-K106 locus for the Papua-New-Guinean population, where the observed value is 0.069 and the expected is 0.128.

To quantify the amount of genetic diversity that occurs between populations, the F_{ST} statistic was calculated using the PopGene statistical package, for each of the loci (Tables 4.7 to 4.10). The highest value obtained, 0.0988, was for the HERV-K115 locus which indicates that 90.12 % of the genetic variation of this locus is within a population (Table 4.7). The next highest value, 0.0696, was for the HERV-K113 locus which signifies that 93.04 % of the genetic variability is in a population (Table 4.8). The third successive value was 0.0303 for the HERV-K103 locus, where 96.97 % of the genetic diversity is within a population (Table 4.9). The lowest value obtained was 0.0063 for the HERV-K106 locus which subsequently implies that 99.37 % of the genetic variation of this locus is within a population (Table 4.10).

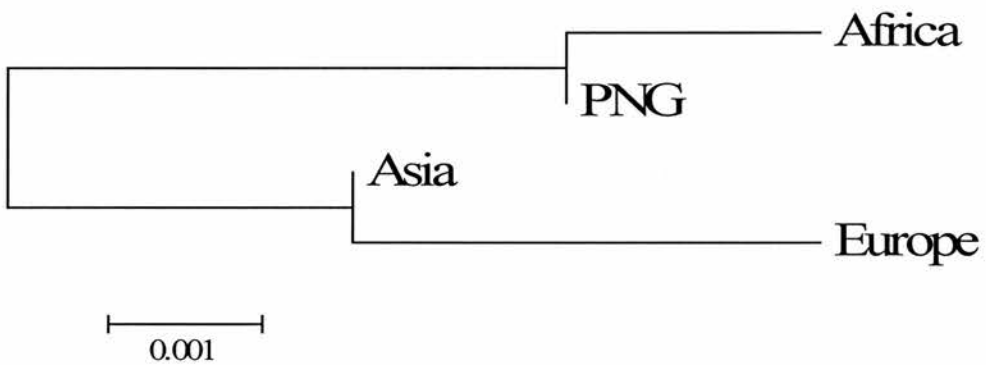
Unbiased genetic distance estimates between each of the population groupings for all loci were calculated using the PopGene statistical software package which utilises the algorithm by Nei, (1978) (Table 4.15). The genetic distance between the African and European populations is the highest at 0.0103, and is followed by; the Papua-New-Guineans and Europeans at 0.0083; the Africans and Asians at 0.0061; the Papua-New-Guineans and Asians at 0.0029; the Europeans and

Asians at 0.0013; and finally the Africans and Papua-New-Guineans at 0.0007. These distances imply that the African and Papua-New-Guinean populations are more closely related to each other than any other population (Figure 4.15).

Table 4.15 Genetic Distances between Africans, Asians, Europeans, and Papua-New-Guineans by Analysis of Variable HERV-K(HML-2) loci

	Africa	Asia	Europe
Asia	0.0061	--	--
Europe	0.0103	0.0013	--
Papua-New-Guinea	0.0007	0.0029	0.0083

Figure 4.15 Genetic distances between Africans, Asians, Europeans, and Papua-New-Guineans presented as a Neighbour-Joining tree.



4.3 Discussion

In order to determine the validity of variable HERV-K loci as markers for examining the dispersal of contemporary humans, the human genomic diversity of five human specific HERV-K(HML-2) proviral loci, which were variable within the human genome databases, was determined by the PCR screening of 109 individuals from four geographical populations. Computational screening for allelic variation within 122 HERV-K loci showed that all the variable loci belonged to the HERV-K(HML-2) subfamily and were human specific. This indicates that the intra-element recombination events that lead to the formation of structural HERV-K loci variation are occurring in quick succession following proviral integration (Macfarlane and Simmonds, 2004).

As computational screening within the human genome databases for allelic variance could be subject to ascertainment bias (Myers et al., 2002), two monomorphic loci, HERV-K107 (M14123) and HERV-K 3q27.2 (AC069420), were also PCR screened for potential novel alleles within all individuals. These two loci were selected on the basis of their apparent recent insertion within the human genome, which was determined by the accumulative nucleotide difference of their respective LTRs of which HERV-K107 manifested 2 or 4 differences and HERV-K 3q27.2 three differences (Section 3.2.1). Each of these loci was subsequently determined to be monomorphic for a complete provirus (refer to Appendix B, Figure B.4), in accordance with the human genome databases, indicating that for human specific HERV-K(HML-2) insertions the human genome databases reflect human genomic diversity.

The microevolution of any DNA sequence within contemporary human populations is influenced by four evolutionary forces which can cause changes in allele frequencies over time; mutation; natural selection; genetic drift; and gene flow. As all loci and populations examined within this study were observed to fit within the Hardy-Weinberg Equilibrium, all populations are presumed to be randomly mating and the loci not subject to selection, this in turn determines that the variable alleles are neutral markers. Subsequently, the major evolutionary forces that could affect the genetic diversity of the HERV-K(HML-2) loci will be; mutation; gene flow and drift.

Previous analysis of the allele frequencies of the insertionally polymorphic loci, HERV-K113 and HERV-K115, within 23 geographically dispersed individuals indicated that their worldwide frequency was 0.19 and 0.04 respectively (Turner et al., 2001). In this study of the variance in 109 geographically dispersed individuals, the allele frequencies were determined to be 0.138 and 0.0734 respectively. The difference of the HERV-K113 frequency can be attributed to a bias towards an African sample within the study by (Turner et al., 2001) and likewise, the higher frequency of the HERV-K115 provirus within this study can be attributed to the inclusion of a representative Sahulian population. Former analysis of the distribution of the HERV-K108 tandem repeat in 27 geographically dispersed individuals (Reus et al., 2001), indicated that 25 (92.5 %) possessed at least one copy of the tandem repeat, suggesting that on a worldwide scale the allele had almost reached fixation. In direct contrast, this study of the variance within 91 individuals indicated that 46 (50.5 %) possessed at least one copy of the tandem repeat. These dissimilar results are highly significant when considering the genetic variability of the HERV-K108 locus

and emphasise the requirement of a large sample size within any population based study. In addition, the predicted reciprocal product of intra-element recombination which led to the generation of the tandemly repeated provirus, was also detected within this study, proving the aberrant intra-element recombination mechanism which generated the novel tandem repeat. As only one individual out of 91 screened possessed this solitary LTR in a heterozygous state, this implies that only one in 200 individuals will possess this allele, accentuating the view that a large number of individuals need to be screened in order to detect less frequent novel HERV-K alleles.

With the exception of the HERV-K113 provirus, all allele frequencies for the novel HERV-K alleles were highest within the African sample. This is exemplified by the presence of the solitary HERV-K103 and HERV-K108 LTR alleles. Similarly, levels of genetic diversity (heterozygosity) were also highest in the African population. Generally, it could be considered that the higher diversity reflects the greater antiquity of the African population, whereby the non-African populations are younger hence less diverse, in compliance with the 'Out Of Africa' model of modern human origins (Cann et al., 1987; Tishkoff et al., 1996; Harris and Hey, 1999). However, it is also possible that the greater African diversity could be a reflection of a larger long-term effective population size (Relethford and Harpending, 1994; Stoneking et al., 1997; Relethford and Jorde, 1999), in keeping with both the 'Multiregional' and 'Out of Africa' models. In these scenarios the apparent containment of the HERV-K103 solitary LTR within the African population, could either be because it has arisen relatively recently or that it has never left the African continent through gene flow or dispersal.

Levels of heterozygosity are expected to be reduced in isolated populations through drift whereby gene flow, which maintains high levels of heterozygosity, is restricted. For the HERV-K113 and HERV-K106 loci, the Papua-New-Guinean sample produced significantly deviant results for the observed and expected levels of heterozygosity. The HERV-K113 locus had an observed heterozygosity of 0.461 and an expected heterozygosity of 0.355. In contrast, the HERV-K106 locus had an observed heterozygosity of 0.069 and expected heterozygosity of 0.128. These results can be interpreted as a representation of a founder population, whereby the majority of individuals possessed the HERV-K113 provirus. Consecutively, the low level of observed heterozygosity of the HERV-K106 solitary LTR, is a reflection of drift suggesting a low level of gene flow, hence that the Papua-New-Guinean population has remained relatively genetically isolated since initial colonisation. Similar levels of heterozygosity and their implications to the history of the Papua-New-Guineans has also been documented for the alpha-globin region (Roberts-Thomson et al., 1996) and mtDNA control region (Redd and Stoneking, 1999).

Measure of the magnitude of the genetic difference between the populations determined within this study was estimated by the application of Wright's F_{ST} statistic. In accordance with other genetic studies (Cavalli-Sforza et al., 1994; Batzer et al., 1994; Jorde et al., 1995), the results indicated that contemporary humans are a very homogenous species with 90.12 % to 99.37 % of genetic variation at HERV-K(HML-2) loci being within a population. Both the 'Out of Africa' and 'Multiregional' models of human expansion, predict that contemporary populations will be largely homogenous, either as a result of resent arisal and dispersal or through worldwide gene flow.

In order to determine the pattern of genetic distance between the four populations and to further determine which evolutionary model the HERV-K(HML-2) loci reflected, Nei's unbiased estimate of genetic distance (Nei, 1978) was calculated for all loci and populations. Subsequently in accordance with other genetic studies (Jorde et al., 2000; Yu et al., 2002; Watkins et al., 2003), the African population appeared to be the most genetically distant and was most similar to the Papua-New-Guinean population. Moreover, the European and Asian populations were more similar than either was to the African population. The close relationship of the Papua-New-Guinean and African populations can be construed to reflect the 'Multiregional' model whereby a high level of gene flow has been maintained between the two populations. However, as the heterozygosity levels of the HERV-K113 and HERV-K106 loci suggest that the Papua-New-Guinean population has remained relatively isolated, the 'Out of Africa' model is the more viable hypothesis. The close relationship of the European and Asian populations and their reduced heterozygosities are consistent with a bottleneck or a successive series of bottleneck events that have reduced the genetic diversity among non-African populations (Harpending and Rogers, 2000). This pattern is not compatible with either the 'Out of Africa' (regional replacement) or 'Multiregional' models as it does not imply either the single large population expansion from Africa or extensive worldwide gene flow.

If the 'Out of Africa' model is assumed to be correct, then it is presumed that there has been more gene flow out of Africa than into Africa. This in theory would lead to reduced levels of heterozygosity within Africa as the population was subject to drift. However, this cannot be the case as the African sample displays both the

highest overall heterozygosity and genetic variability. This can be accounted for if Africa has retained a large long-term effective population size, or if gene flow back into Africa has occurred, consistent with the Multiregional Model.

It would appear that variable HERV-K(HML-2) loci do serve as valid markers for examining the dispersal of contemporary human populations. Although they do not endorse either the 'Out of Africa' or 'Multiregional' model of human evolution, they reflect that their genetic history has been more complicated than either model. Such a history has also been inferred from the population based analysis of coding and non-coding autosomal regions (Harding et al., 1997; Zhao et al., 2000; Yu et al., 2001). The statistical data generated from the frequency of the allelic variants of the HERV-K(HML-2) loci, implies a combination of the presumptions of both models is likely to be the case. This is consistent with an 'Assimilation Model' of human dispersal, whereby the traits associated with contemporary human populations arose in Africa but they were subject to a complex and potentially long-term process of interbreeding and population movement.

CHAPTER 5

HERV-K AS FACILITATORS OF CHROMOSOMAL REARRANGEMTS

5.1 Introduction

Studies aimed towards determining the qualitative and quantitative genetic differences between humans and the great apes have primarily focused upon examining cytogenetic and single base-pair differences. These have revealed that the ancestral human karyotype has been subject to two major rearrangements; the first the fission of an ancestral chromosome to produce human chromosomes 14 and 15 which is present in all members of the superfamily hominoidea (Haig, 1999); and the second the fusion of two ancestral chromosomes to form human chromosome 2, which occurred after the divergence of humans and chimpanzees (Yunis and Prakash, 1982). At the single nucleotide level, human and chimpanzee have been determined to be 98.4 to 98.8 % identical in both coding and non-coding DNA regions (Kaessmann et al., 2001; Chen and Li, 2001; Fujiyama et al., 2002), suggesting that the regions encompassing these differences may be of special relevance to understanding the divergence of the two species (King and Wilson, 1975; Navarro and Barton, 2003). However, recent large-scale comparative analysis of the human genome to homologous regions within the great apes has shown that both large (10 to 400 kb) and small scale (< 10 kb) insertions, deletions, duplications, inversions and translocations are also of primary importance to understanding the source of biological differences between the species (Fujiyama et al., 2002; Frazer et al., 2003; Locke et al., 2003; Tsend-Ayush et al., 2004). Such chromosomal rearrangements may have played a role in the reproductive isolation of the hominoids (Gagneux and Varki, 2001).

Large scale differences (10 to 400 kb) are associated with regions termed low-copy repeats (LCRs) which are estimated to constitute at least 5 % of the human genome. HERV sequences represented by complete proviruses are associated with this classification of repetitive DNA as they are ~ 10 kb in length (Stankiewicz and Lupski, 2002b). Also included are paralogous segmental duplications which require 95 to 97 % similarity in order to serve as substrates for non-allelic recombination (Stankiewicz and Lupski, 2002a). Comparative studies of the genomes of closely related primates indicate that LCR regions have been targets of rapid evolutionary turnover (Eichler, 2001). Successive duplication has led to the expansion and adaptive evolution of the Morpheus and Olfactory gene families during the divergence of hominoid apes (Johnson et al., 2001; Eichler et al., 2001; Trask et al., 1998; Newman and Trask, 2003).

Small scale differences (< 10 kb) between the genomes of the great apes are primarily associated with highly repetitive sequences of which transposable elements constitute the highest proportion. In addition to acting as insertional mutagens, interspersed repetitive elements such as SINEs and LINEs have also been observed to serve as substrates for homologous recombination. A LINE-LINE recombination event mediated the inversion of human chromosomal regions, Yp and Xq21, following the divergence of human and chimp (Schwartz et al., 1998) and *Alu-Alu* (SINE) recombination events have been observed to lead to several human diseases (Deininger and Batzer, 1999; Kolomietz et al., 2002). Interestingly, a third of LCR regions within the human genome contain an enrichment of *Alu*s, indicating that spread of genomic duplications may have been facilitated by *Alu* elements (Babcock et al., 2003; Bailey et al., 2003). Conversely, analysis of genomic rearrangements

which have taken place after the divergence of human and chimpanzee on chromosome 21, show that no one specific class of repetitive element is prevalent at the boundaries of segmental deletions and insertions (Frazer et al., 2003).

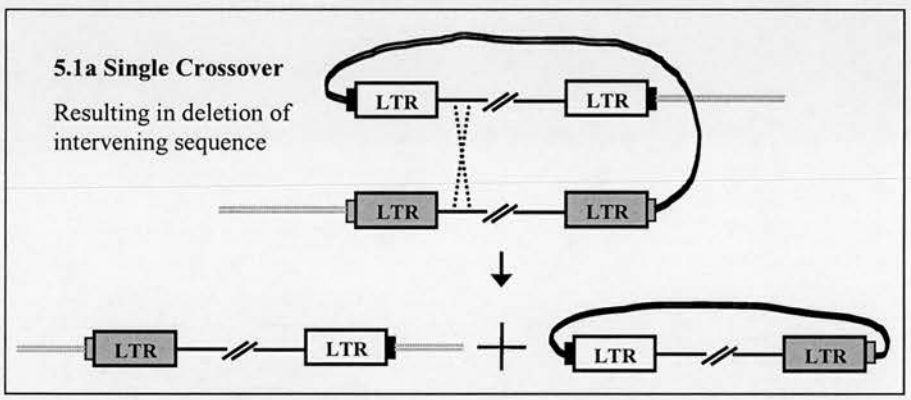
With the exception of non-allelic recombination between two HERV-15 proviruses on the Y chromosome (Blanco et al., 2000; Kamp et al., 2000; Sun et al., 2000; Bosch and Jobling, 2003), there is no evidence that inter-element recombination between HERVs is a frequent cause of structural mutation within humans. The high diversity of each subfamily and the low copy number of HERV proviruses within the human genome limits the availability of compatible HERV sequences (Babcock et al., 2003). Despite this, HERV sequences are highly recombingenic (Hughes and Coffin, 2001) with at least 16 % of HERV-K(HML-2) proviral sequences being estimated to have been involved in non-allelic recombination during primate evolution (Johnson and Coffin, 1999) and HERV-16 sequences being suggested to have played a major role in the duplication of PERB11 and HLAcI genes (Kulski et al., 1999a; Kulski et al., 1999b). The recently re-characterised family of retrotransposon, SVA (SINE, VNTR and Alu), is derived from SINE.R and HERV-K(HML-2) elements (Zhu et al., 1994; Ostertag et al., 2003) and a chimeric HERV-H / HERV-K retroelement transposed onto chromosomes 10, 19 and Y before the divergence of the Homininae. In addition, the copy number of the HERV-W family has increased within the human genome via retroviral transposition and LINE mediated retrotransposition (Pavlicek et al., 2002). Recombination or gene conversion has also led to the concerted evolution of the HERV-H family and (Mager and Freeman, 1995) and resulted in the homogenisation

of the LTRs of the RTVL-1a, HERV-K110 / K18 proviral loci (Johnson and Coffin, 1999) and ERV-K(C4) region in higher primates (Dangel et al., 1995).

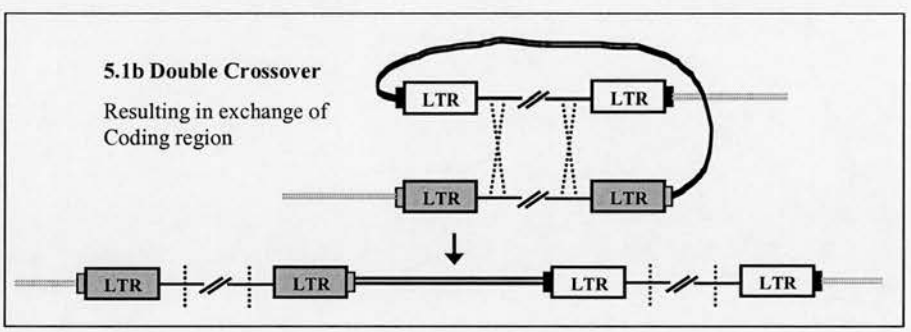
Recombination between both allelic and non-allelic HERV proviral sequences is expected to result in chromosomal rearrangements which manifest as insertions, deletions, duplications, inversions and translocations (Figures 5.1, 5.2 and 5.3). Aberrant events which affect the germ line arise as a germinal mutation in one of the parents of an affected individual and assuming they are non-deleterious, can subsequently be passed onto the next generation.

Deletion of DNA sequence can either arise as a result of inter-element recombination between two HERVs which are located on the same chromosome (Figure 5.1a) or through intra-element recombination whereby only the structural regions of a provirus are lost and a solitary LTR is generated (Section 4.2.4, Figure 4.7). Loss of intervening DNA on the X or autosomal chromosomes is expected to be detrimental to an organism as the deletion will generate phenotype imbalance. Nevertheless, presuming that a deletion on the Y chromosome does not affect a region required for subsistence, the deletion can be passed onto further generations as only one copy of the Y chromosome is required within an individual (Blanco et al., 2000). In addition to deletion, allelic recombination between sister chromatids which contain a HERV provirus is also expected to result in duplication (Figure 5.2a) and is exemplified by the polymorphic HERV-K108 locus in humans (Section 4.2.5). Although duplication will also create phenotypic imbalance, it has been a major driving force of evolutionary change amongst the hominoids (Fortna et al., 2004). Non-allelic recombination between HERV proviruses located on different chromosomes is expected to result in the translocation of chromosomal segments (Figure 5.3a)

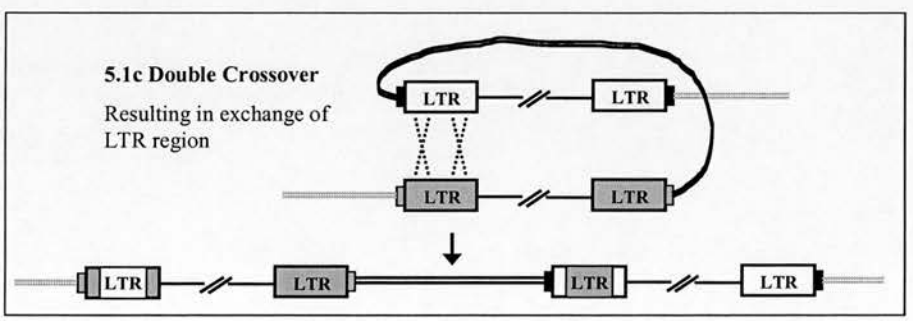
Figure 5.1 Outcome of Homologous Recombination between HERVs located on the same Chromosome.



LTRs and Direct Repeats Divergent

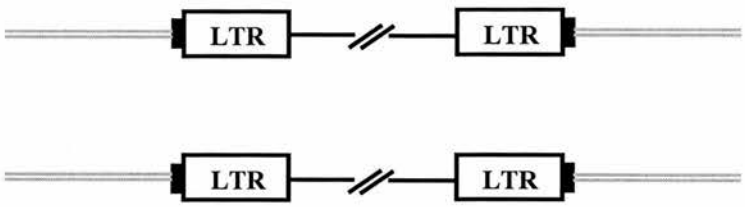


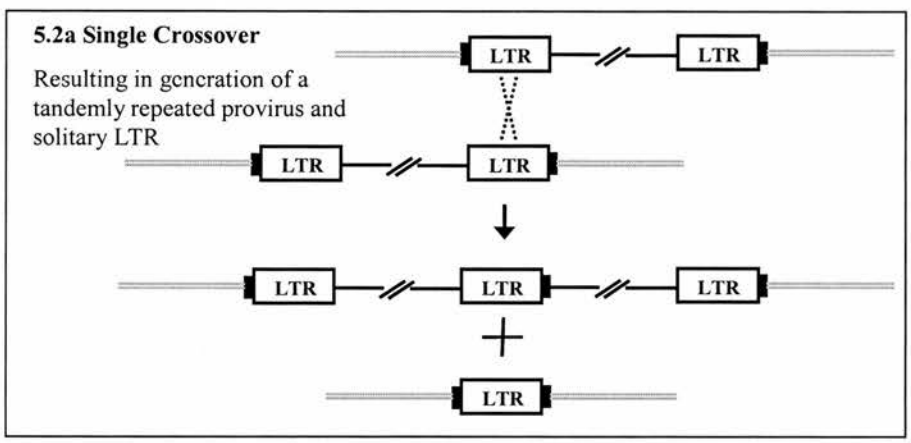
LTRs and Direct Repeats not Divergent



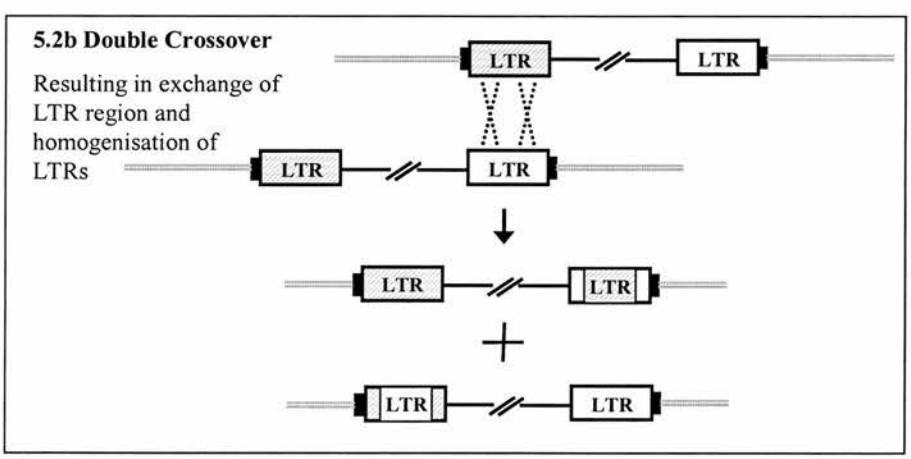
LTRs Divergent Direct Repeats the same

Figure 5.2 Outcome of Homologous Recombination between Sister Chromatids containing a HERV Provirus.



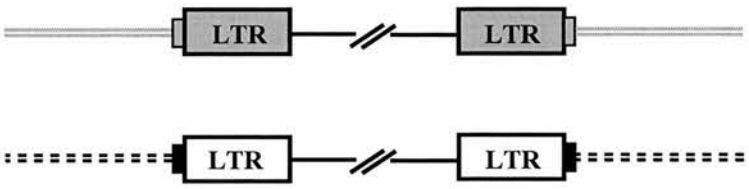


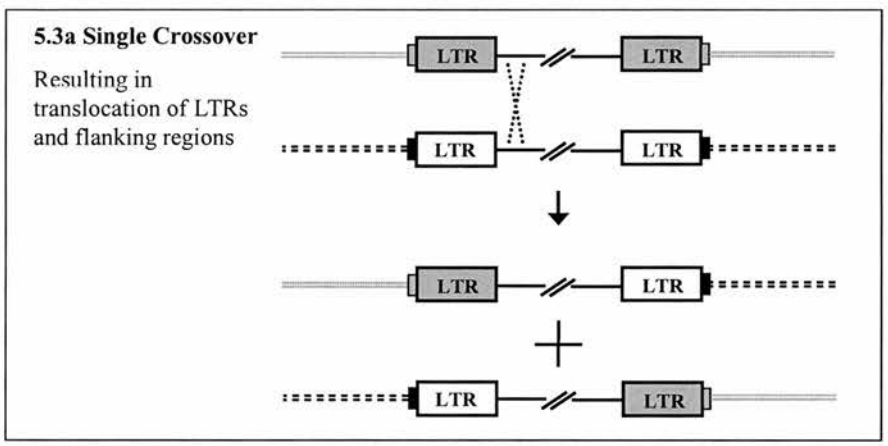
LTRs and Direct Repeats not Divergent

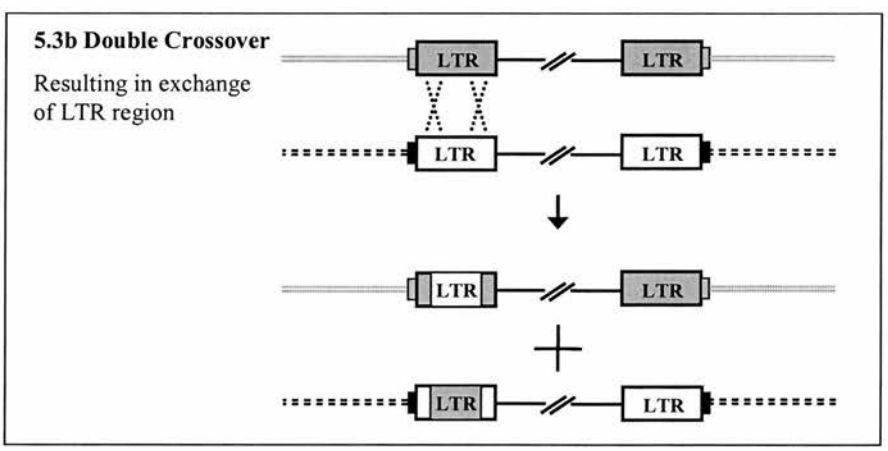


LTRs Divergent Direct Repeats the same

Figure 5.3 Outcome of Homologous Recombination between HERV Proviruses located on different chromosomes.







(Johnson and Coffin, 1999).

If a such an event occurs within the germ cells of an individual, it is expected that a quarter of their offspring will be genetically normal, a quarter will be balanced carriers for the translocation and the remaining 50 % of offspring will be non-viable as they carry incompatible combinations of the translocated chromosomal regions. This implies that a translocation event between autosomes has a one in four chance of becoming part of the gene pool.

Recombination events which have led to sequence exchange within or between HERV proviruses can be identified via the construction of phylogenetic trees utilising the LTRs of proviral sequences (Johnson and Coffin, 1999; Hughes and Coffin, 2001). As reverse transcription generates two identical LTRs during the retrotransposition of a provirus, over time the LTRs are expected to remain more similar to each other than to the LTRs of any other HERV integration (Johnson and Coffin, 1999). However, if a proviral sequence has undergone sequence exchange as a result of either inter-element recombination (Figures 5.1a and 5.3a) or non-allelic gene conversion (Figure 5.1c and 5.3b), it is expected that the LTRs will be highly divergent and not cluster together in a phylogenetic tree. Alternatively, sequence exchange between sister chromatids can result in the homogenisation of LTRs (Figure 5.2b) which will write over the signature of former sequence exchange or mutations accumulated over time. As previously described, sequence exchange between sister chromatids can result in the duplication and reciprocal deletion of a proviral sequence (Figure 5.2a); in such circumstances all respective LTRs will cluster within a phylogenetic tree as they are derived from the same proviral locus (Section 4.2.1, Figure 4.1). Similarly, intra-element recombination between the 5'

and 3' LTRs of an individual provirus will result in the production of a solitary LTR which will also group with the progenitor proviral LTRs (Section 4.2.1, Figure 4.1). Allelic or non-allelic sequence exchange involving regions other than the LTRs (Figure 5.1b) will not be detectable within a phylogenetic tree constructed from LTR sequences unless a translocation event is involved (Figures 5.1b and 5.3a).

In common with infection by exogenous retroviruses, the retrotransposition and integration of a HERV results in the generation of short target site duplications of 4 to 6 bp at the integration site (Macfarlane and Simmonds, 2004). These direct repeat sequences flank either end of the newly integrated provirus and are identical at the time of integration. HERV-K(HML-2) proviruses which possess disparate target site duplications have been suggested to be the end product of inter-element homologous recombination (Hughes and Coffin, 2001). Non-allelic recombination events which are likely to produce HERV sequences with variable direct repeats involve either proviruses located on the same chromosome (Figure 5.1a), or proviruses situated on different chromosomes (Figure 5.3a). The outcome of such sequence exchange is likely to dramatically alter karyotype with the first event resulting in deletion and the second translocation of chromosomal regions.

To examine the structural mutagenic impact of HERVs upon the human genome, the direct repeats of 108 unique HERV-K insertions from three different subgroups were compared. Fourteen loci were subsequently verified to possess disparate target site duplications, of which eight appeared to variable as a result of inter-element recombination. Phylogenetic analysis of the LTRs belonging to 47 complete HERV-K proviruses, was then used to investigate whether any originated through non-allelic sequence exchange. The events that led to the inconsistent target

site duplications of two human specific HERV-K(HML-2) insertions were then determined by comparative analysis to the pre-integration site in non-human primates. The results showed that disparate target site duplications can be generated by mechanisms other than inter-element recombination.

5.2 Results

5.2.1 Comparison of Direct Repeats

The target site duplications of 108 HERV-K insertions from three different subgroups and of variable relative age were compared to examine the potential impact of HERV-K sequences in the remodelling of the catarrhine genome (Table 5.1). The three near-complete HERV-K(HML-2) human specific proviruses; 12q24.11 (AC002350), HERV-K(C19) (AF017229), and 21q21.1 (AL109763), along with the six human specific solitary LTRs; 2q21.2 (AC084028), 11q13.3 (AP0001184), 12q13.3a (AC079034), 14q23.3 (AL139022), 16p13.12 (AC009167), and 16q23.1 (AC009132) were not included in the data set as they had lost one or more of their direct repeats (Tables 5.2 and 5.3). The HERV-K(HML-2) LTRs, 2 (AC027778) and 6 (AL157379), which are present in both chimpanzee and human, were also not included as they contained incomplete LTRs (Table 5.3). In addition, the Type I HERV-K(HML-2) near-complete provirus, at 10q24.2 (AL392107) was excluded as the last 634 bp of the 5'LTR was absent (Table 5.2). Of the 13 HERV-K(HML-3) proviruses, all possessed intact 5' and 3' LTRs (Table 5.4 and Appendix B, Table B.4). Finally, only a single HERV-K(HML-4) provirus, 10p15.1 (AL391427), was excluded from the data set as 826 bp of the beginning of the 3'LTR was absent (Table 5.5). In total, this left; 28 HERV-K(HML-2) proviruses, 61 HERV-K(HML-2) solitary LTRs, 13 HERV-K(HML-3) proviruses, and 6 HERV-K(HML-4) proviruses that were suitable for further analysis.

Table 5.1 Comparison of the Target Site Duplications of 108 unique HERV-K insertions within the Human Genome

Type of HERV-K element screened	Total no of unique insertions examined	No. of insertions with variable direct repeats due to nucleotide substitution	No. of insertions with variable direct repeats due to potential inter-element homologous recombination
HML-2 Human specific provirus	15	0	0
HML-2 provirus present in Human and Chimp	1	1	0
HML-2 provirus present in Human, Chimp and Gorilla	6	1	0
HML-2 provirus present in Human, Chimp, Gorilla and Orang	2	0	2
HML-2 provirus present in Human, Chimp, Gorilla, Orang and Gibbon	2	1	1
HML-2 novel provirus reported in Macfarlane and Simmonds (2004)	2	1	0
HML-2 Human specific solitary LTRs	49	1	2
HML-2 solitary LTRs present in Human and Chimp	7	0	0
HML-2 solitary LTRs present in Human, Chimp and Gorilla	5	0	1
HML-3 provirus	13	0	2
HML-4 provirus	6	1	0

Table 5.2 Inconsistent Direct Repeat sequences of HERV-K(HML-2) Proviruses. [a] Relative age determined by the presence or absence of the provirus in primate species (Section 3.2.3). Branching dates from humans; 55 Mya for Prosimians, 45 Mya for New World Monkeys, 28 for Old World Monkeys, 20 Mya for Gibbons, 14 Mya for Orang-utans, 8.5 Mya for Gorillas and 6.3 Mya for Chimpanzee (Figure 3.8). Variable - Disparate target site duplications. U – Relative age undetermined. Δ 3' LTR - Part or all of the 3'LTR and direct repeat is deleted within the human genome. Δ 5' LTR - The 5'LTR and direct repeat is deleted within the human genome. NA - Not applicable. Variations in single nucleotides are underlined. To view HERV-K(HML-2) proviruses with consistent direct repeat sequences, refer to Appendix B, Table B.2.

(HML-2) Provirus	Age [a] (Mya)	Accession	5'Direct Repeat	3'Direct Repeat	Features
3p25	< 20	AC018829	CTTGGT	GAAAGT	Variable
19p13.11	< 20	AC011467	TCCCAG	TGTAAT	Variable
6p22.1	< 45	AL121932	GATCCC	CCTGGG	Variable
4q32.3	< 8.5	AC106872	CTTTCT	TTTTAT	Variable
6p21.1	< 45	AL035587	AAACT	AAATT	Variable
19q13.13	< 28	AC012309	GGTCTT	GATCTT	Variable
Xq28	U	(Human) AF277315	CCAGC	CCACC	Variable
		(Chimp) AC144385	CCAGC	CCACC	Variable
21q21.1	< 6.3	AL109763	GCCAGG	NA	Δ 3' LTR
HERV-K(C19)	< 6.3	AF017229	NA	AGGTAT	Δ 5' LTR
12q24.2	< 6.3	AC002350	GTATT	NA	Δ 3' LTR
10q24.2	U	AL392107	NA	GGTGC	Δ 5' LTR

Table 5.3 Inconsistent Direct Repeat sequences of HERV-K(HML-2) Solitary LTRs. [a] Relative age determined by the presence or absence of the LTR in primate species (Section 3.2.2). Branching dates from humans; 55 Mya for Prosimians, 45 Mya for New World Monkeys, 28 for Old World Monkeys, 20 Mya for Gibbons, 14 Mya for Orang-utans, 8.5 Mya for Gorillas and 6.3 Mya for Chimpanzee (Figure 3.8). Variable - Disparate target site duplications. Δ LTR - The solitary LTR is partially deleted within the human genome. NA - Not applicable. Variations in single nucleotides are underlined. To view HERV-K(HML-2) Solitary LTRs with consistent direct repeat sequences, refer to Appendix B, Table B.3.

HML-2 Solitary LTR	Age [a] (Mya)	Accession	5'Direct Repeat	3'Direct Repeat	Features
7p21.2	< 6.3	AC006035	AGGCAA	GGATGA	Variable
17q22	< 6.3	AC032016	ACAAAT	ACGATT	Variable
	< 6.3				
5q35.3	< 6.3	AC023559	GATAAA	GATACA	Variable
Xq13.1	< 14	AJ239320	CTCTG	AGTCA	Variable
2q21.2	< 6.3	AC084028	NA	NA	Δ LTR
11q13.3	< 6.3	AP001184	CAACT	NA	Δ LTR
12q13.3a	< 6.3	AC079034	CAAGAA	NA	Δ LTR
14q23.3	< 6.3	AL139022	NA	NA	Δ LTR
16p13.12	< 6.3	AC009167	NA	NA	Δ LTR
16q23.1	< 6.3	AC009132	NA	NA	Δ LTR
2	< 8.5	AC027778	NA	NA	Δ LTR
6	< 8.5	AL157379	AGAGG	NA	Δ LTR

Table 5.4 Inconsistent Direct Repeat sequences of HERV-K(HML-3) Proviruses. Variable - Disparate target site duplications. To view HERV-K(HML-3) proviruses with consistent direct repeat sequences, refer to Appendix B, Table B.4.

HML-3 Provirus	Accession	5'Direct Repeat	3'Direct Repeat	Features
5q14.3	AC117524	GTAACC	TGATAA	Variable
19p13.11	AC010615	CTTACA	GTATAA	Variable
7q21.3	(Human)AC069292 (Chimp) AC142300	TTTTCA TTTTCA	TTTTCA TTTTCA	

Table 5.5 Inconsistent Direct Repeat sequences of HERV-K(HML-4) Proviruses. Δ 3' LTR - Part or all of the 3'LTR and direct repeat is deleted within the human genome. Variable - Disparate target site duplications. Variations in single nucleotides are underlined. To view HERV-K(HML-4) proviruses with consistent direct repeat sequences, refer to Appendix B, Table B.5.

HML-4 Provirus	Accession	5'Direct Repeat	3'Direct Repeat	Features
17q21.31	AC109326	G<u>C</u>TTC	G<u>G</u>CTC	Variable
10p15.1	AL391427	ACTATA	NA	Δ 3'LTR

The four HERV-K(HML-2) proviruses; 4q32.3 (AC106872), 6p21.1 (AL035587), 19q13.13 (AC012309), and Xq28 (AF277315) all had different direct repeats which appeared to be as a result of nucleotide substitution (Tables 5.1 and 5.2). Therefore differences between each of the respective target site duplications were unlikely to be a by-product of inter-element recombination. Interestingly, the chimpanzee ortholog (AC114385) of the HERV-K(HML-2) provirus situated at Xq28 (AF277315), possessed exactly the same incongruent direct repeats as the human version of the provirus (Table 5.2).

The three remaining HERV-K(HML-2) proviruses that possessed disparate target site duplications were of considerable relative age (Tables 5.1 and 5.2). The proviruses located at 3p25 (AC018829) and 19p13.11 (AC011467) within the human genome, are also present in all other members of the family Hominidae (Section 3.2.3, Table 3.10). The provirus situated in the human genomic region 6p22.1 (AL121932) also has orthologs in all members of the super-family Hominoidea (Section 3.2.3, Table 3.10).

Of the human specific HERV-K(HML-2) solitary LTRs, the element at 5q35.3 (AC023559) appeared to possess disparate target site duplications as a result of substitution, and the two LTRs at 7p21.1 (AC006035) and 17q22 (AC032016), showed incongruent direct repeats suggestive of inter-element recombination (Table 5.3). The direct repeats of the HERV-K(HMl-2) LTR at Xq13.1 (AJ239320), which has orthologs in chimpanzee and gorilla, were also consistent with inter-element recombination (Table 5.3).

Two of the 13 HERV-K(HML-3) proviruses located at, 5q14.3 (AC117524) and 19p13.11 (AC010615), each possessed dissimilar target site duplications

indicative of inter-element recombination (Table 5.4). None of the remaining 11 HERV-K(HML-3) proviruses possessed variable direct repeats (Appendix B, Table B.4).

From the total of six HERV-K(HML-4) proviral insertions, only one element located at 17q21.31 (AC0109326) had disparate target site duplications (Table 5.5 Appendix B, Table B.5). As two point mutations at the second and third nucleotides within one of the duplications could account for the incongruence, it appears that this element has not undergone inter-element recombination.

In summary, comparative sequence analysis of the target site duplications of 108 unique HERV-K insertions indicated that 6 out of 108 (5.55 %) varied as a result of substitution. The inconsistency of the direct repeats belonging to a further 8 insertions, suggests that 7.4 % of HERV-K sequences have been involved in non-allelic recombination during the evolution of the human genome. Interestingly, potential chromosomal rearrangements facilitated by HERV-K proviruses are found in elements of both considerable relative age and of relatively recent insertion. This suggests that chromosome restructuring events may have occurred throughout primate evolution.

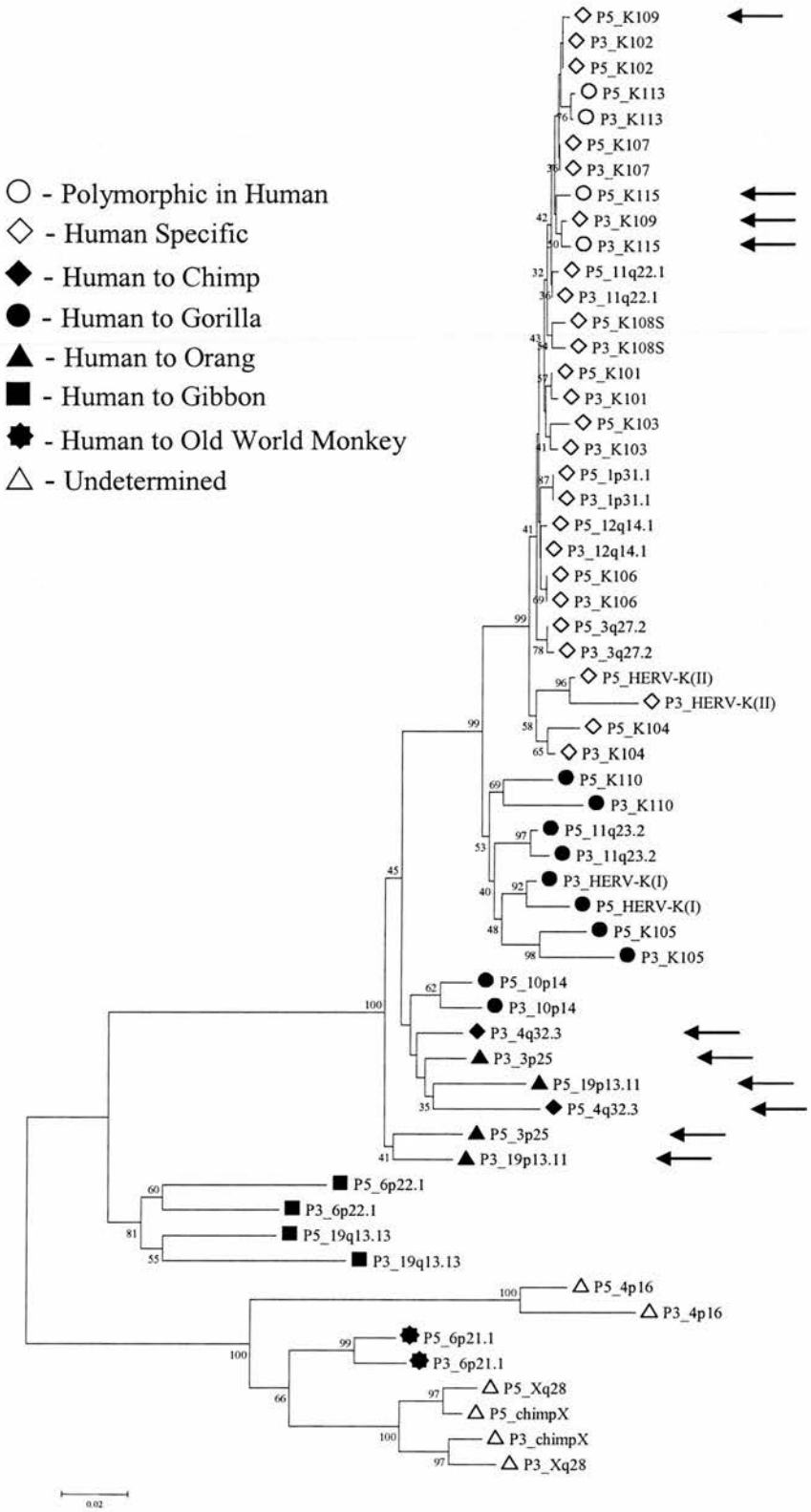
5.2.2 Topology of HERV Proviral LTRs in Phylogenetic Trees

If sequence exchange in the form of inter-element recombination has occurred between non-allelic HERV sequences, in addition to the exchange of flanking sequence, the LTRs of each provirus would be expected to be divergent in a plylogenetic tree (Section 5.1, Figures 5.1a and 5.3a). Sequence exchange between LTR regions will also produce a similar result, but the expected corollary is unlikely to affect the target site duplications or flanking region sequence (Section 5.1, Figures 5.1c and 5.3b). To further examine genetic relationships of HERV proviruses and the potential role of inter-element recombination, neighbour-joining trees were constructed for the three HERV-K subgroups examined in Section 5.2.1, using the Kimura-2-parameter distance estimate with alignment gaps being handled as a complete deletion and the phylogeny tested with 500 bootstrap replications.

For the HERV-K(HML-2) subgroup, the 5' and 3' LTRs of 28 proviruses of variable relative age were aligned by hand using the SIMMONIC sequence analysis package (Appendix A, Alignment A.2) and a neighbour-joining tree constructed using MEGA, version 2.1. The near-complete proviruses; 12q24.11, HERV-K(C19), 21q21.1, and 10q24.2 were excluded from the data set as all or part of one of their LTRs is absent within the human genome sequence databases (Section 5.2.1, Table 5.2).

One human specific, HERV-K109, and one insertionally polymorphic provirus, HERV-K115, possessed LTRs which were divergent within the phylogenetic tree (Figure 5.4). As each of these unique integrations had previously been determined to possess identical target site duplications (Section 5.2.1, Table

Figure 5.4 Phylogeny of LTRs belonging to Complete HERV-K(HML-2) Proviruses. The neighbor-joining tree is based on the Kimura-2-parameter distance estimate. Bootstrap values above 30 % out of 500 re-samplings are shown at the internodes. The arrows highlight the 5' and 3' LTRs that do not cluster.



5.2), the non-clustering LTRs can be presumed to reflect gene conversion. This view is further confirmed by the number of nucleotide differences between the LTRs of the HERV-K115 provirus (Section 3.2.3, Table 3.10). Assuming the rate of accumulative mutation affecting HERV-K proviruses is constant, the 14 nucleotide differences of the insertionally polymorphic HERV-K115 provirus was higher than found in monomorphic human specific proviruses (Section 4.3).

The 5' and 3' LTRs of the HERV-K(HML-2) provirus located at 4q32.3, which integrated within the common ancestor of human and chimpanzee also deviated from expectation (Figure 5.4). Although the direct repeats of this provirus were determined to be variable as a result of two nucleotide substitutions (Section 5.2.1, Table 5.2), lack of LTR clustering suggests that the LTR regions have undergone sequence exchange. Interestingly, the LTRs of this provirus also clustered with deviant LTRs belonging to two older proviruses, HERV-K 3p25 and HERV-K 19p13.11. Both of these proviruses also possessed highly variable target site duplications (Section 5.2.1, Table 5.2). In addition, all three of these proviruses possessed exactly the same sequence deletion of 1937bp within the *pol* region, further indicating that they form a very distinctive group of HERV-K(HML-2) proviruses (Section 3.2.1, Table 3.5).

The target site duplications of the HERV-K(HML-2) proviruses situated at 6p21.1 and 19q13.13, were previously determined to be disparate due to nucleotide substitution (Section 5.2.1, Table 5.2). This view is further supported by the grouping of their LTRs and the respective bootstrap values of 99 % and 55 % (Figure 5.4).

High bootstrap values of 97 % support the association of the 5' and 3' LTRs of the human and chimp orthologs of the HERV-K(HML-2) provirus situated within

the human pseudoautosomal region of the X chromosome (Xq28) (Figure 5.4). This denotes that since the presence of this element within the common ancestor of chimp and human, neither homologue has undergone sequence exchange across the LTR regions.

Interestingly, the 5' and 3' LTRs of the HERV-K(HML-2) provirus located at 6p22.1 clustered with a bootstrap value of 60 % (Figure 5.4), although previous analysis of the target site duplications indicated that this element may have undergone inter-element recombination (Section 5.2.1, Table 5.2). This insinuates that either the direct repeats vary as a result of a mechanism other than inter-element recombination or presuming that inter-element recombination has occurred, the LTRs have since been homogenised. However, as the bootstrap value is less than 90 %, there can be little confidence in the clustering of the HERV-K 6p22.1 LTRs.

Overall, the positioning of the unique HERV-K(HML-2) proviral insertions within the phylogenetic tree supports their relative evolutionary age with bootstrap values ranging from < 30% to 100 %, in reasonable support of the division of major lineages (Figure 5.4). However, low bootstrap values were obtained for the association of LTRs belonging to elements of greater relative age, which could be due to representative sample bias. This would entail that older 'intermediate' proviral insertions are no longer embodied as complete proviruses within the human genome, leading to a predisposition towards the more intact recent integrations. This is exemplified by the number and division of human specific elements at the top of the tree (Figure 5.4). With the exception of the five insertions; HERV-K115, HERV-K109, HERV-K 4q32.3, HERV-K 3p25, and HERV-K 19p13.11, all HERV-K(HML-2) proviral LTRs cluster, indicative of a lack of sequence exchange. In total,

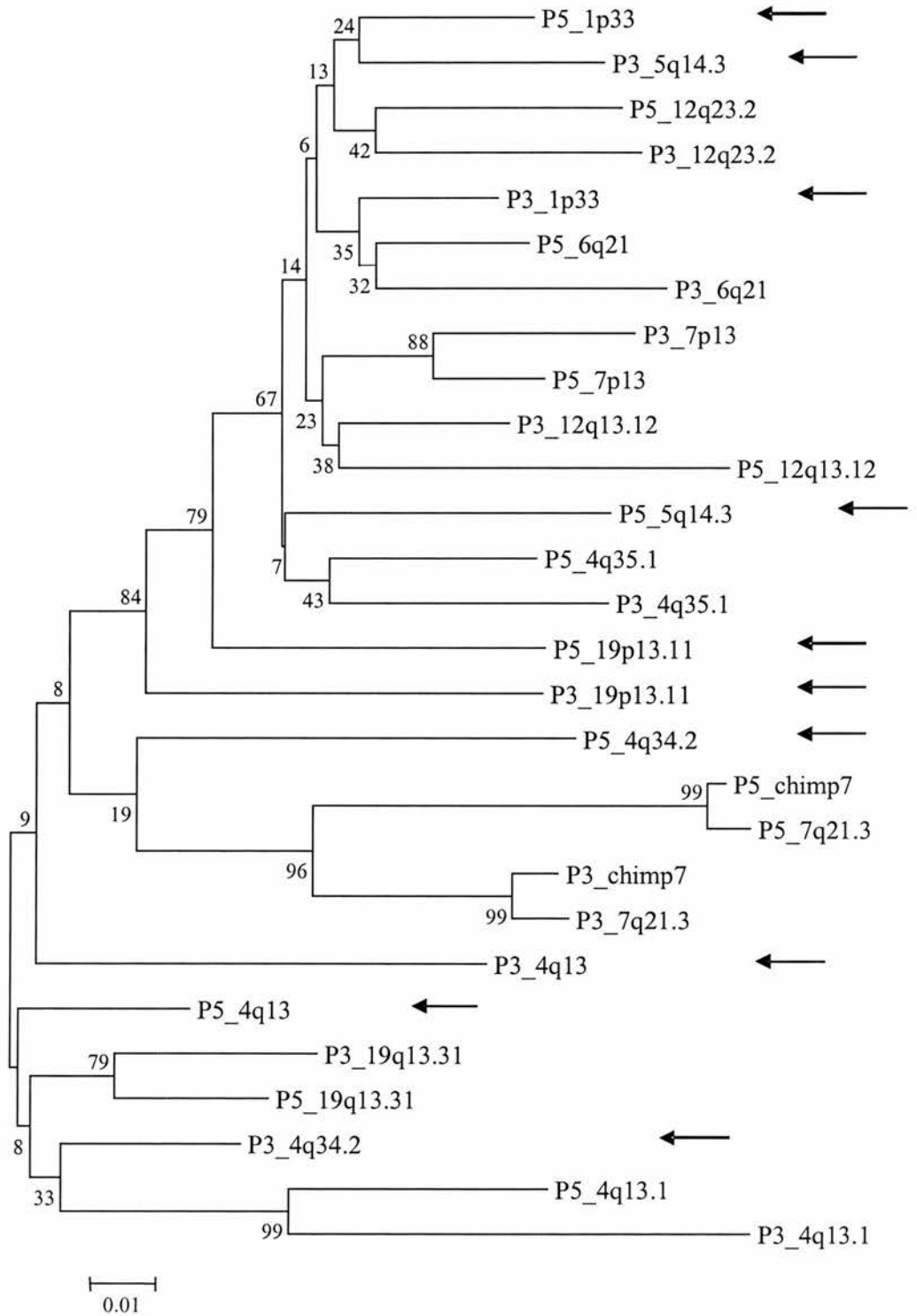
two out of 23 HERV-K(HML-2) proviruses appear to have been involved in non-allelic recombination events.

To analyse the HERV-K(HML-3) subgroup, the 5' and 3' LTRs of 13 proviruses were aligned by hand using the SIMMONIC sequence analysis package (Appendix A, Alignment A.5) and a neighbor-joining tree constructed using MEGA, version 2.1. The respective LTRs of the four proviruses located at; 7p13, 7q21.3, 19p13.31, and 4q13.1 had high bootstrap values of 79 % to 99 % at the internode, in support of their grouping (Figure 5.5). The four proviruses situated at; 12q23.2, 6q21, 12q13.12, and 4q35.1 also possessed LTRs which clustered within the tree. However, the bootstrap values at each of the internodes ranged from 32 % to 43%, suggesting a low degree of confidence in their grouping within the phylogenetic tree. Similarly, bootstrap values at the common internodes for the respective LTRs of the proviruses located in the cytogentic regions; 1p33, 4q34.2, and 4q13 were also low. As the respective target site duplications of these three proviruses were previously determined to be identical (Section 5.2.1 and Appendix B, Table B.4), and the divergent LTRs were supported by low bootstrap values (Figure 5.5), it cannot be concluded that the LTR regions have been subject to gene conversion.

High bootstrap values of 99 % support the association of the 5' and 3' LTRs of the human and chimp orthologs of the HERV-K(HML-3) provirus located within the human cytogenetic region 7q21.3 (Figure 5.5). This denotes that since the presence of this element within the common ancestor of chimp and human, neither homologue has undergone sequence exchange across the LTR regions.

The two remaining HERV-K(HML-3) proviruses, HERV-K 5q14.3 and HERV-K 19p13.11, were previously determined to possess disparate target site

Figure 5.5 Phylogeny of LTRs belonging to Complete HERV-K(HML-3) Proviruses. The neighbor-joining tree is based on the Kimura-2-parameter distance estimate. Bootstrap values out of 500 re-samplings are shown at the internodes. The arrows highlight the 5' and 3' LTRs that do not cluster.

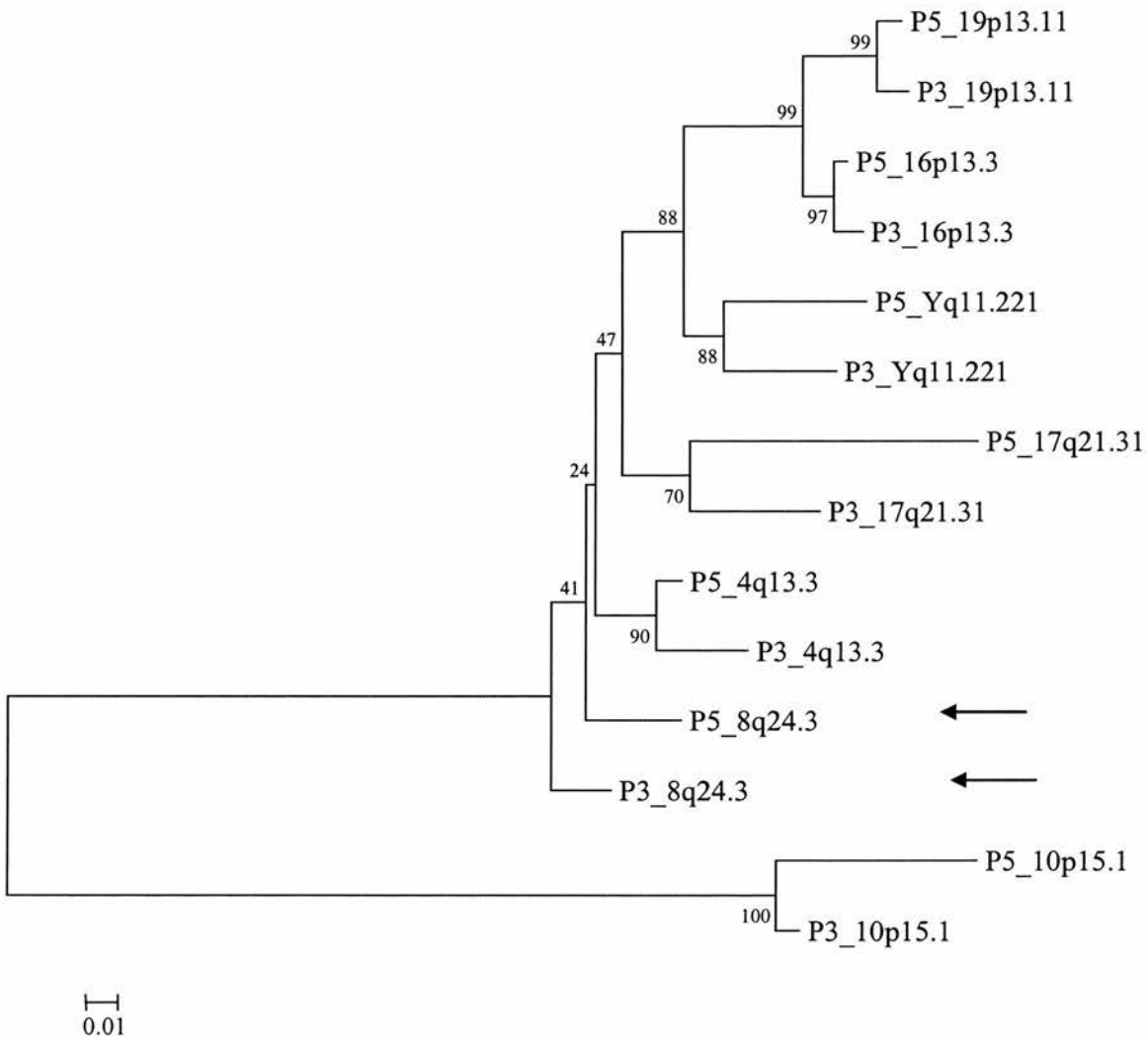


duplications indicative of inter-element recombination (Section 5.2.1, Table 5.4). High bootstrap values of 67 % and 84 % support the divergence of their respective LTRs (Figure 5.5), providing additional evidence that these loci have been involved in non-allelic recombination.

In order to analyse the HERV-K(HML-4) subgroup, the 5' and 3' LTRs of six proviruses were aligned by hand using the SIMMONIC sequence analysis package (Appendix A, Alignment A.7) and a neighbor-joining tree constructed using MEGA, version 2.1. The near-complete provirus located at 10p15.1 was excluded from the data set as part of the 3' LTR was absent within the human genome sequence databases (Section 5.2.1, Table 5.5). With the exclusion of the provirus situated at 8q24.3, all of the respective LTRs belonging to the HERV-K(HML-4) proviruses had high internode bootstrap values which ranged from 99 % to 100 %, in support of their association (Figure 5.6). This implies a lack of sequence exchange across the LTR regions of these five proviral genomes. The target site duplications of the HERV-K(HML-4) provirus situated at 17q21.31 were previously determined to be disparate as a result of nucleotide substitution (Section 5.2.1, Table 5.5). The bootstrap value of 100 % at the internode for the respective 5' and 3' LTR supports this view.

Of the total of 47 complete HERV-K proviruses considered, four possessed both disparate target site duplications and highly divergent LTRs. This implies that 8.51 % of the HERV-K proviruses examined may have been involved in non-allelic recombination. It must be considered that with the exception of two orthologous chimpanzee sequences, all sequence data examined is from the human genome sequence databases and so does not reflect the full evolutionary history of the

Figure 5.6 Phylogeny of LTRs belonging to Complete HERV-K(HML-4) Proviruses. The neighbor-joining tree is based on the Kimura-2-parameter distance estimate. Bootstrap values above 30 % out of 500 re-samplings are shown at the internodes.



HERV-K proviral insertions within non-human primates. Therefore it is possible that orthologs of the HERV-K insertions analysed here may have facilitated non-allelic recombination in the genomes non-human primates, following the evolutionary divergence of each of the species. In addition, the construction of phylogenetic trees using proviral LTRs may be flawed due to the phenomenon of intra-element LTR homogenisation. If homogenisation occurred following an inter-element recombination event, evidence of sequence divergence would be obscured.

5.2.3 Analysis of Pre-Integration Sites and Flanking Regions of Atypical HERV-K

If HERV-K sequences with inconsistent direct repeats and highly divergent LTR regions are the end product of homlogous recombination between non-allelic proviral sequences, then a reciprocal HERV-K element with an opposite configuration of direct repeats and flanking region would also be expected to be generated (Section 5.1, Figures 5.1a and 5.3a). From the total of 108 unique HERV-K insertions examined in Section 5.2.1, eight appeared to have disparate target site duplications indicative of inter-element recombination, of which four were also determined to also possess highly divergent LTRs (Section 5.2.2, Figures 5.4 and 5.5). To further examine the possibility that these 'hybrid' sequences were a product of inter-element recombination, the human genome databases were screened for each of the expected reciprocal products; none of the predicted sequences were present (Table 5.6). This implies that either the reciprocal products are not present in the representative individuals screened by the human genome project or that the expected reciprocal sequences do not form a constituent of the contemporary human gene pool.

In order to investigate the relative age of the events that generated the disparate the target site duplications of the HERV-K(HML-2) proviral sequences contained at 3p25 and 19p13.11, the flanking regions of each insertion were amplified in chimpanzee and gorilla. Unique 5' and 3' flanking region primers were designed according to the human genome sequences AC018829 and AC011467 and

Table 5.6 HERV-K elements which possess Diagnostic Signatures of Inter-element Recombination or Gene Conversion. [a] The direct repeats appeared to be variable as a result of nucleotide substitution and low bootstrap values support LTR divergence. [b] The direct repeats were the same and low bootstrap values support LTR divergence. NA – Not applicable.

HERV-K element	Disparate Direct Repeats	Divergent LTRs	Reciprocal Present in Human Genome
HML-2 Proviruses			
19p13.11	Yes	Yes	No
3p25	Yes	Yes	No
6p22.1	Yes	No	No
4q32.3 [a]	Yes	Yes	No
K115	No	Yes	NA
K109 [b]	No	Yes	NA
HML-2 Solitary LTRs			
7p21.2	Yes	NA	No
17q22	Yes	NA	No
Xq13.1	Yes	NA	No
HML-3 Proviruses			
5q14.3	Yes	Yes	No
19p13.11	Yes	Yes	No
1p33 [b]	No	Yes	NA
4q34.2 [b]	No	Yes	NA
4q13 [b]	No	Yes	NA

screened by standard nucleotide-nucleotide BLAST against the non-redundant and high-throughput sequence databases, to ensure that DNA sequences were unique (all primer sequences and combinations are shown in Section 2.3.2, Table 2.4). The direct repeat disparity of both proviruses was subsequently determined to be analogous in all three species, indicating that the potential aberrant events occurred relatively quickly following integration (Figure 5.7).

To examine the timing of potential changes in karyotype which are insinuated by the disparate site duplications of the HERV-K(HML-2) solitary LTR at Xp13.1, which has orthologs in chimpanzee and gorilla, unique 5' and 3' flanking region primers were designed according to the human genome sequence AJ239320. The primers were then screened by standard nucleotide-nucleotide BLAST against the non-redundant and high-throughput sequence databases, to ensure that DNA sequences were unique (all primer sequences and combinations are shown in Section 2.3.2, Table 2.4).

Under the presumption that the solitary LTR at Xp13.1 may have undergone inter-element recombination at any point during the evolutionary divergence of the homininae, variability of the flanking regions within the chimpanzee and gorilla orthologs, would be the expectation The corollary of such circumstances would be a negative PCR result when amplifying the entire solitary LTR in one of the non-human primates using primers specific to the human ortholog. To confirm the HERV-K integration was present in all three species, PCR amplification was performed using the unique 5' flanking region primer and universal HERV-K(HML-2) LTR antisense primer. As expected all three primates produced an amplicon of predicted size. PCR amplification was then performed for the entire solitary LTR

Figure 5.7 Sequence Alignment of the Flanking Regions of the HERV-K(HML-2) Proviruses located at 3p25 and 19p13.11.

Figure 5.7a Sequence Alignment of the 5' and 3' Flanking Regions of the HERV-K(HML-2) Provirus located at 3p25 in Human, which is also present in Chimpanzee, Gorilla and Orang-utan. Alignment gaps are indicated by dashes (-). The direct repeats are highlighted in bold and underlined.

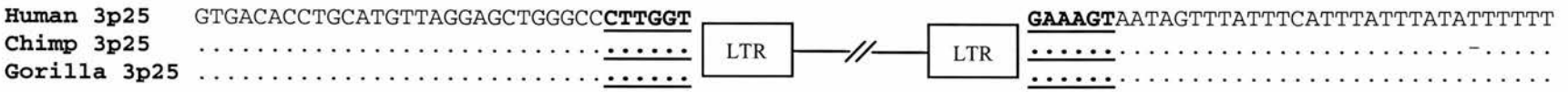
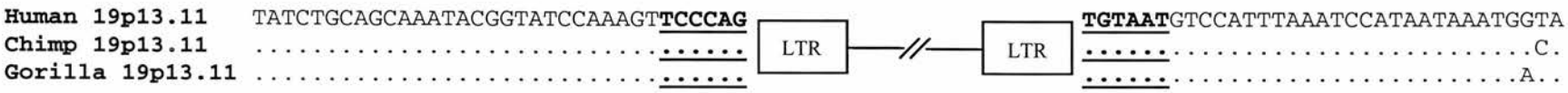


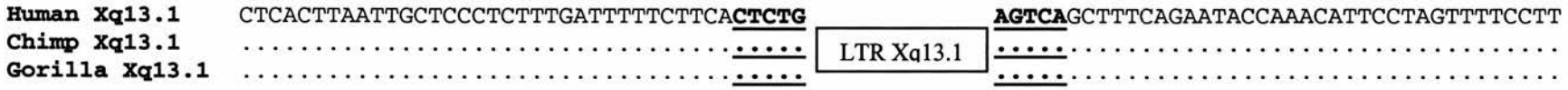
Figure 5.7b Sequence Alignment of the 5' and 3' Flanking Regions of the HERV-K(HML-2) Provirus located at 19p13.11 in Human, which is also present in Chimpanzee, Gorilla and Orang-utan. Nucleotide substitutions at each position are indicated with the appropriate nucleotide. The direct repeats are highlighted in bold and underlined.



using the human 5' and 3' flanking region primers. Contrary to expectation, all three species produced amplicons which were equivalent in size. Sequence analysis of the three amplicons revealed that all species shared the same flanking regions, solitary LTR and direct repeat incongruity (Figure 5.8). This implied that the direct repeat incongruity arose relatively quickly following the integration of the element as it must have been present in the common ancestor of all three species.

In order to confirm that more recent HERV-K(HML-2) insertions with variable direct repeats were a product of inter-element recombination, unique 5' and 3' flanking region primers were designed for the human specific solitary LTRs at 7p21.2 (AC006035) and 17q22 (AC032016). The primers were then screened by standard nucleotide-nucleotide BLAST against the non-redundant and high-throughput sequence databases, to ensure that DNA sequences were unique. As the screening of potential primers revealed that the flanking regions of each insertion were highly repetitive, two 5' primers were designed for each solitary LTR to enable hemi-nested amplification reactions to be performed (all primer sequences and combinations are shown in Section 2.3.2, Table 2.4) The conformational PCR amplification for each of the solitary LTRs, verified the absence of each solitary LTR in non-human primates. This dictated that the HERV-K insertions were either not fixed in the gene pool at the time of gorilla / chimpanzee / human divergence or that they integrated during hominid evolution. Following the presumption that the disparate direct repeats of the human specific solitary LTRs were a signature of inter-element recombination, amplification for the pre-integration site in non-human primates was performed under the expectation that amplicons would not be produced. Contrary to expectation distinct PCR products were generated, suggesting

Figure 5.8 Sequence Alignment of the 5' and 3' Flanking Regions of three HERV-K(HML-2) insertions with Disparate Direct repeats in Human, Chimpanzee and Gorilla. The direct repeats are highlighted in bold and underlined.



that the disparate target site duplications were not a signature of inter-element recombination (Figures 5.9 and 5.10).

Intriguingly, the amplicons corresponding to the pre-integration site of the solitary LTR at 7p21.2 were 250 bp shorter than expected (Figure 5.9). Sequence analysis of the PCR products with comparison to the orthologous region in human revealed that the human specific solitary LTR was flanked by a 250 bp duplication of target site sequence (Figure 5.11a). Interestingly, a similar scenario was also observed for the near-complete human specific HERV-K(HML-2) provirus at 21q21.1 (AL109763) whereby the 3' LTR appears to be truncated by a sequence paralogous to the 5' flanking sequence (Figure 5.11b). As it can be assumed that upon entry the provirus at 21q21.1 was composed of two identical LTRs, this duplication must have occurred following integration.

Sequence analysis of the human specific solitary LTR at 17q22 and respective pre-integration site in non-human primates indicated that the downstream direct repeat was 4 bp shorter than the upstream (Figure 5.12). Assuming the former state of the solitary LTR to be a complete provirus, the short deletion either occurred following the integration of the HERV-K sequence or during integration when perhaps an incomplete target site duplication of 2 bp was generated.

Analysis of the flanking regions and pre-integration sites of three HERV-K(HML-2) solitary LTRs of variable relative age indicates that the events which lead to incongruent target site duplications are occurring relatively quickly following integration. This insinuates that a high degree of homology is required between sequences in order for sequence exchange to occur. This further implies that rearrangements are unlikely to be facilitated by highly divergent HERV-K

Figure 5.9 Amplification for the Pre-Integration Site of the Human specific HERV-K(HML-2) Solitary LTR at 7p21.2 (AC006035) in Chimpanzee and Gorilla. The Lanes entitled; 'H' contains human DNA, 'Ch' chimpanzee DNA, 'Gor' Gorilla DNA and (-ve) is a negative DNA control. Presuming the disparate direct repeats were a result of homologous recombination, the pre-integration site would not be expected to amplify.

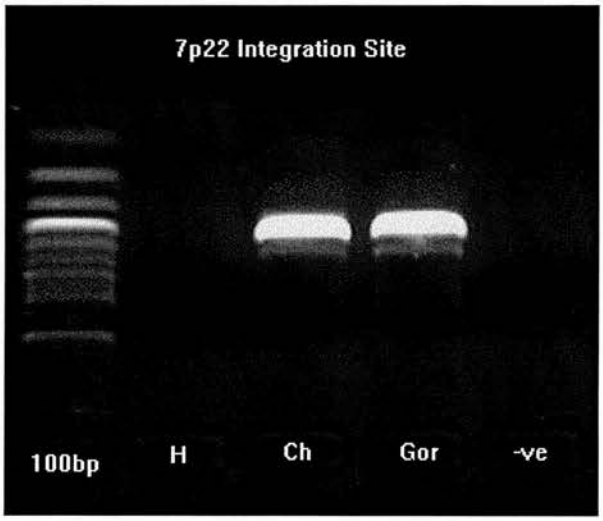


Figure 5.10 Amplification for the Pre-Integration Site of the Human specific HERV-K(HML-2) Solitary LTR at 17q22 (AC032016) in Chimpanzee and Gorilla. The Lanes entitled; 'H' contains human DNA, 'Ch' chimpanzee DNA, 'Gor' Gorilla DNA and (-ve) is a negative DNA control. Presuming the disparate direct repeats were a result of homologous recombination, the pre-integration site would not be expected to amplify.

Figure 5.11 Flanking Region Duplication leading to Variable Direct Repeat Sequences.

Figure 5.11a HERV-K(HML-2) Solitary LTR at 7p21.2. The pre-integration sequence in chimpanzee and gorilla is represented by the top figure and contains a 250bp sequence with the respective nucleotide sequences GGATGA and AGGCAA at each end. The human specific solitary LTR at 7p21.2 (Accession AC006035) has inconsistent direct repeat sequences AGGCAA and GGATGA which are underlined. Sequence comparison indicates that the solitary LTR is flanked by a 250bp duplication of the pre-integration sequence.

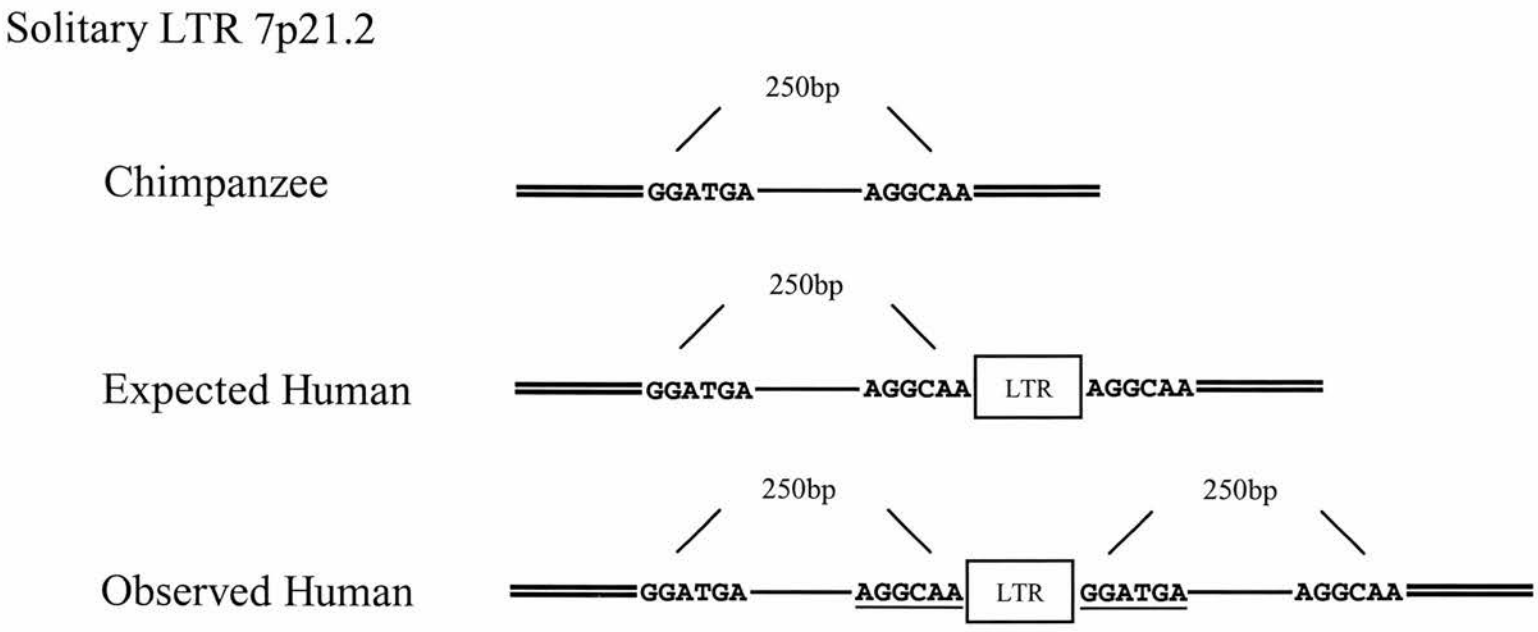


Figure 5.11b HERV-K(HML-2) Near Complete Provirus at 21q21.1. The first figure represents the pre-integration sequence present in chimpanzee (Accession BS000043). The near complete human specific provirus at 21p21.1 (Accession AL109763) contains a truncated 3'LTR of 257bp which is adjacent to a 450bp duplication of the pre-integration site sequence.

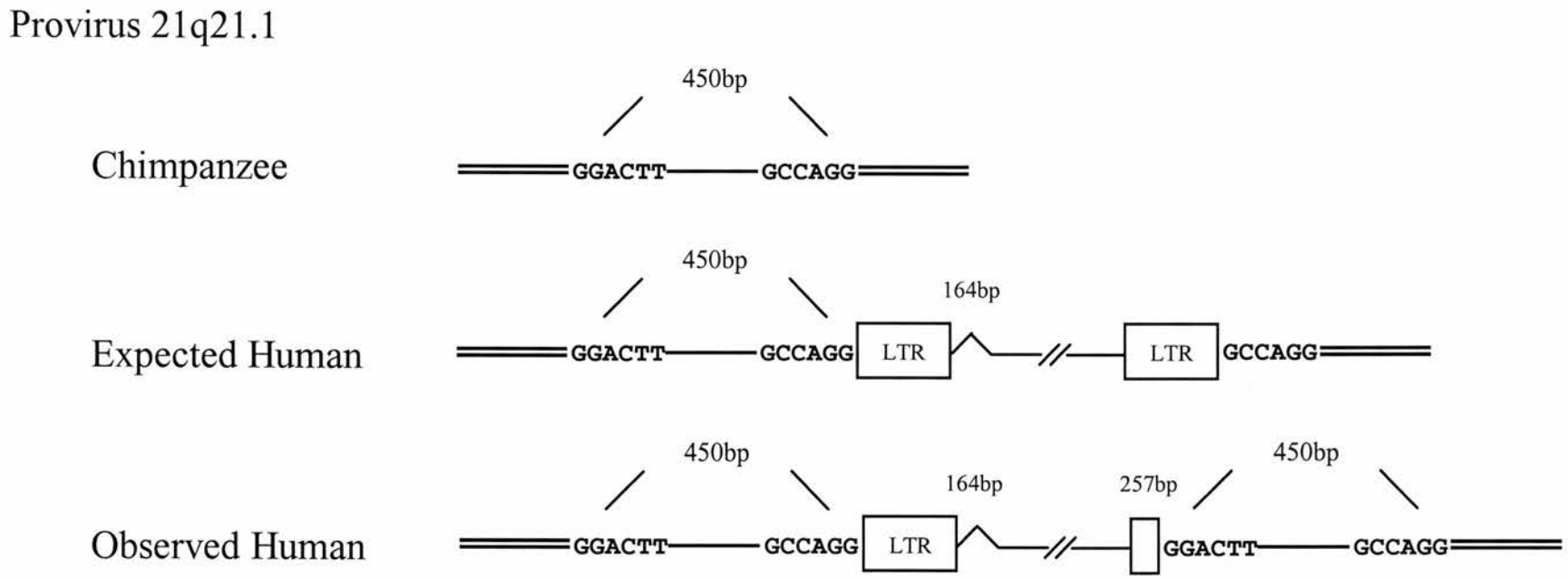


Figure 5.12 Sequence Alignment of the Human specific HERV-K(HML-2) Solitary LTR at Chromosomal Location 17q22 and Corresponding Pre-Integration Site in Non-Human Primates. The human sequence of the HERV-K(HML-2) LTR containing locus (Accession AC032016) is shown on the top line. Nucleotide substitutions at each position are indicated with the appropriate nucleotide. Alignment gaps indicated by dashes (-). The direct repeats of the solitary LTR are highlighted in bold and underlined.

```
Human       CTACAAGGTTATTAATTGCAACACTTTTTATAATA ACAAAT  LTR 17q22  ACGATT TAGAATATCTTTTTTGTATTATACTTTAAGTTTT
Chimpanzee  .....G.....................................----------.....................................
Gorilla     .....G.....................................----------.....................................
```

sequences, for example between those of different relative age. Of importance is the observation that disparate target site duplications can be generated by mechanisms other than inter-element recombination. Consequently, the total number of elements estimated to have been involved in non-allelic recombination will be overestimated when conducting analysis as described in Sections 5.2.1 and 5.2.2. Furthermore, the impact of inter-element recombination between HERV-K sequences to the remodelling of cattarrhine genomes will also be overestimated.

5.3 Discussion

Previous analysis of sequence exchange between HERV-K(HML-2) loci, suggests that at least 16 % of proviral sequences contained within the human genome are an end product of inter-element recombination, which is expected to have resulted in large-scale chromosomal rearrangements (Hughes and Coffin, 2001). This figure was reached as 6 out of 35 proviral insertions examined were observed to be flanked by inconsistent target site duplications. In this study, which considered 108 insertions belonging to three different HERV-K subfamilies, 14 proviral sequences (12.96 %) were determined to possess disparate direct repeats. Six of these loci were established to be variable as a result of nucleotide substitution, whereby one or two point mutations in one of the direct repeats could account for the disparity. As nucleotide substitution was not considered in Hughes and Coffin, (2001), the assessment of the number of elements involved in inter-element recombination is likely to be an overestimate. This is exemplified by the HERV-K(HML-2) provirus at 4q32.3 which is deemed to have disparate direct repeats as a result of substitution within this study, but is regarded as an end product of non-allelic recombination in Hughes and Coffin, (2001). Interestingly, within this study the chimpanzee ortholog of the HERV-K(HML-2) provirus contained at Xq28 within the human genome, possessed the same direct repeat inconsistency as the human sequence, indicating that the point mutation arose within a common ancestor of both species.

Of the remaining loci which possessed variable target site duplications; three were complete HERV-K(HML-2) proviruses, two were HERV-K(HML-3) proviruses and three were HERV-K(HML-2) solitary LTRs. With the exception of

the HERV-K(HML-2) provirus located at 6p22.1, phylogenetic analysis of the LTRs belonging to the complete proviruses indicated that they were highly divergent, providing further evidence that they were an end product of large-scale chromosomal rearrangements.

In order to investigate the relative age of the events that generated the disparate the target site duplications of the HERV-K(HML-2) proviral sequences contained at 3p25 and 19p13.11, the flanking regions of each insertion was amplified in chimpanzee and gorilla. The direct repeat disparity of both proviruses was subsequently determined to be analogous in all species. Similar results were also obtained for the HERV-K(HML-2) solitary LTR at Xp13.1 which integrated within the common ancestor of human, chimpanzee and gorilla. This implies that if these three unique HERV-K(HML-2) insertions are an end product of inter-element recombination, then sequence exchange occurred quickly following integration. Consequently, a high degree of sequence homology might be the requirement for aberrant events to take place, suggesting that sequence exchange will not occur between highly diverse HERV-K elements. A high degree of sequence homology has also been observed to be a requirement for allelic or non-allelic sequence exchange between parologous segmental duplications (Stankiewicz and Lupski, 2002a) and *Alu* retrotransposons (Batzer and Deininger, 2002; Waldman and Liskay, 1988).

To further examine the effect that inter-element recombination has had specifically upon the plasticity of the hominid genome, amplification was conducted for both the pre-integration sites and presence of the two remaining HERV-K(HML-2) solitary LTRs which possessed variable target site duplications. Contrary to expectation, amplification of the pre-integration site in non-human primates was

achieved using primers which flanked the human specific solitary LTRs. This implied that the inconsistent target site duplications of both loci were not an end product of inter-element recombination, which is expected to result in either large scale deletion or translocation of chromosomal DNA. Sequence analysis of the human specific solitary LTR at 17p22 indicated that the 3' direct repeat had undergone deletion resulting in the loss of 4 bp. The most parsimonious explanation for the disparate direct repeats of the human specific solitary LTR at 7p21.2 is that unequal crossover occurred between allelic sequences; one of which contained the HERV sequence and a second which consisted of the pre-integration site. Such an intra-chromosomal recombination event is expected to result in the duplication of the pre-integration site sequence (Macfarlane and Simmonds, 2004). As a second human specific HERV-K(HML-2) proviral insertion was also observed to possess a similar duplication of pre-integration site sequence, is likely that allelic unequal crossover events have occurred frequently during the genomic evolution of the primates. These results are highly significant when performing comparative analysis of the target the site duplications of any retroelement family. The frequency of inter-element recombination could be grossly overestimated if it is assumed that consistent direct repeats always reflect large-scale chromosomal rearrangements.

It should be considered that the genomic retroviral elements that exist today represent only a small fraction of the total germ line integration and subsequent recombination events that have occurred. If non-allelic recombination takes place between two HERV sequences located upon the same chromosome during meiosis I, the outcome (deletion of intervening DNA) is likely to be detrimental to the host as phenotype imbalance is generated. The exception is inter-element recombination

between sequences located upon the Y chromosome as only one copy of this chromosome is required in order for survival. This is likely to account for the numerous chromosomal rearrangements observed to be present upon the human Y chromosome (Bachtrog, 2003; Hurles and Jobling, 2003; Hurles et al., 2004) and includes those mediated by HERV sequences (Bosch and Jobling, 2003). Likewise, if non-allelic recombination takes place between two HERV sequences located upon different chromosomes during meiosis I, the outcome (translocation of chromosome) will result in 50 % sterility. In order for a translocation event to become part of the gene pool, the translocation must be balanced with the reciprocal outcome. Such an event is expected to occur in one out of four offspring. However, the next generation will also have 50 % sterility rate, further reducing the likelihood of such an event becoming part of the gene pool. Subsequently, if a HERV sequence is an end product of a translocation, the reciprocal sequence should be present in the gene pool; this was not the case for any of the insertions possessing disparate direct repeats in this study. To confirm if the four HERV-K complete proviruses with disparate direct repeats and highly divergent LTRs are an end product of inter-element recombination, the pre-integration site should be examined, as was carried out for the human specific HERV-K(HML-2) solitary LTRs in this study.

Of interest is the HERV-K(HML-2) complete provirus located at 6p22.1 within the human genome. This insertion was determined to possess inconsistent target site duplications which were presumed to have been generated by inter-element recombination. However, phylogenetic analysis of the LTRs indicated that they were not highly divergent, implying that the insertion was not an end product of large-scale chromosomal rearrangement. Two scenarios can explain this incongruity,

the first is that following inter-element recombination the LTRs of the insertion became homogenised through sequence exchange between sister chromatids. The second is that that variable direct repeats were generated by a mechanism other than inter-element recombination. Considering the observations made within this study, the second scenario is more likely to be the case.

Sequence exchange between allelic and non-allelic HERV-K sequences can result in the conversion of internal sequence which does not affect the flanking regions of the insertions. This has resulted in; the homogenisation (Johnson and Coffin, 1999) and accelerated divergence of LTR regions (Turner et al., 2001; Macfarlane and Simmonds, 2004), the concerted evolution of the HERV-H family (Mager and Freeman, 1995), and the exchange of coding region within the HERV-K(HML-2) subfamily (Costas, 2001; Macfarlane and Simmonds, 2004). Such gene conversion events occur frequently between repetitive sequences contained within the human genome (Liao et al., 1998; Pavelitz et al., 1999; Tremblay et al., 2000; Mefford et al., 2001; Roy-Engel et al., 2002; Hurles et al., 2004) and are predicted to arise through gap repair during unequal crossover (Jeffreys et al., 1998; Jeffreys and Neumann, 2002; Jeffreys and May, 2004). In this study chimpanzee orthologs of two HERV-K proviruses were included in the data set to determine the extent of gene conversion. As the respective 5' and 3' LTRs of the orthologous loci clustered, sequence exchange is presumed to have not taken place since the divergence of the two species. The extent of gene conversion will be considered in more detail in Chapter 6.

In conclusion it would appear that HERV-K sequences are highly recombingenic. Divergence of the target site duplications of HERVs can arise as a

result of mutation effecting a few base pairs or local sequence exchange between sister chromatids. In theory, it is highly unlikely that large-scale chromosomal rearrangements mediated by HERV proviruses have been maintained within the primate gene pool as they are expected to be detrimental to the host.

CHAPTER 6

SEQUENCE ANALYSIS OF HERV-K

6.1 Introduction

Retroelements, which rely upon a reverse transcriptase step for their amplification, constitute 90 % of the transposable elements which are present within the human genome (Lander et al., 2001). They are classified into two major groups, the LTR and non-LTR elements. HERVs and their derived retrotransposons are members of the LTR retroelements and make up 8 % of the human genome. Within each of these two major groupings, retroelement families are either autonomous in their replication, whereby they encode all proteins required for their amplification (*cis*), or they are nonautonomous and depend upon another retroelement family to supply the proteins necessary for their replication and integration (*trans*). Of the retroelements which are currently retrotranspositionally active within the human genome, the non-LTR autonomous LINE family are perhaps the best characterised (Sassaman et al., 1997; Boissinot et al., 2000; Sheen et al., 2000; Myers et al., 2002; Salem et al., 2003b; Lutz et al., 2003). The lesser known SVA retrotransposon family, which replicate in *trans* and are derived from the HERV-K subgroup (HML-2), are also known to be currently retrotranspostionally active within human populations (Ono et al., 1987; Zhu et al., 1994; Kim et al., 1999; Ostertag et al., 2003; Bennett et al., 2004).

HERV sequences which are present within the human genome are recognised to have been initially acquired via ancient retroviral infection of germ line cells (Lower et al., 1996; Andersson et al., 1999; Sverdlov, 2000; Tristem, 2000; Benit et al., 2001). Following the original insertion of a provirus, intracellular retrotransposition and recombination have been proposed as the mechanisms by

which particular families have increased in copy number within the germ line (Lower et al., 1996; Patience et al., 1997). Prior to the start of this study, it was generally accepted that the HERV-K family had expanded in copy number following a 'master element' model of retrotransposition, whereby each subgroup had proliferated either in *cis* or *trans*. However, analysis of the topology of HERV-K(HML-2) proviruses and investigation of the selective pressures that had been acting upon HERV-K ORFs, indicated that the history of these elements was far more complex (Zsiros et al., 1998; Zsiros et al., 1999; Costas, 2001). More recently, examination of the *env* region of HERV-K proviruses has implied that this ORF has been subject to purifying selection (Belshaw et al., 2004). This observation is of significance as the *env* region is presumed to be only required for movement between cells. As this is not required within a model of intracellular retrotransposition in *cis* or *trans* it therefore implies that HERVs have proliferated within germ line cells via reinfection.

The catalogue of complete and near complete HERV-K(HML-2), HERV-K(HML-3) and HERV-K(HML-4) proviruses determined within Chapter 3, provides a unique opportunity for examining the selective forces that have acted upon them. Furthermore, as the relative age of HERV-K(HML-2) proviruses was also ascertained, the evolutionary pressures acting at different stages of their proliferation can be determined.

Here, the mode of HERV-K proliferation was considered by analysis of synonymous (dS) and non-synonymous (dN) changes across their ORFs, reconstruction of their phylogeny with analysis of dS/dN between adjacent sequences and comparison to the retrotransposon families LINE and SVA. The results obtained were consistent with those previously reported. However, analysis of the levels

selection acting upon LINE elements demonstrated that expansion in *cis* is a more likely mechanism of HERV-K proliferation. Furthermore, subdivision of HERV-K(HML-2) proviruses into groups of known relative age showed that proviruses of greater relative age were, in the past, under greater purifying selection than those of more recent acquisition. This observation is noteworthy as it suggests that the observed functional constraints of HERV-K ORFs are a remnant of the past and therefore not suggestive of recent purifying selection.

6.2 Results

6.2.1 Mosaic Evolution of HERV-K(HML-2) Proviral genomes

To study the evolutionary relationships of proviruses belonging to the HERV-K(HML-2) subgroup, neighbour-joining phylogenetic trees were constructed using MEGA, version 2.1 (Figures 6.1 to 6.6). The proviral sequences utilised are listed within Appendix B, Table B.6. As the HERV-K(HML-2) proviral LTR sequences were previously observed to have been subject to extensive sequence exchange (Chapters 3, 4 and 5), the LTRs and internal genic regions were considered individually.

Overall the phylogenies were congruent with elements of a similar relative age forming clusters. The exception was the HERV-K(I) provirus, which within the *gag* and *prt* regions grouped with the HERV-K 6p21.1 provirus, an integration of significant greater relative age (Figures 6.3 and 6.4). Interestingly, this association is also reflected within the tree constructed using all internal (non-LTR) regions (Figure 6.2). The grouping of these two proviruses within the *gag* and *prt* regions is likely to be indicative of sequence exchange of short stretches of nucleotides as the two proviruses do not cluster within the LTR, *pol* and *env* topologies.

Superimposition of the Type I and Type II proviral genotypes upon the phylogenetic trees revealed that the Type I proviruses were not monophyletic, as would be expected if the elements have followed a 'master element' model of retrotransposition. As it is assumed that the Type I diagnostic deletion has a single origin, this result is surprising as it does not reflect a clonal expansion of the Type I

Figure 6.1 Phylogeny of LTRs belonging to HERV-K(HML-2) Proviruses, showing the Topology of the Type I and Type II Proviral genomes. The neighbour-joining tree is based on the Kimura-2-parameter distance estimate. Bootstrap values above 30 % out of 500 re-samplings are shown at the internodes.

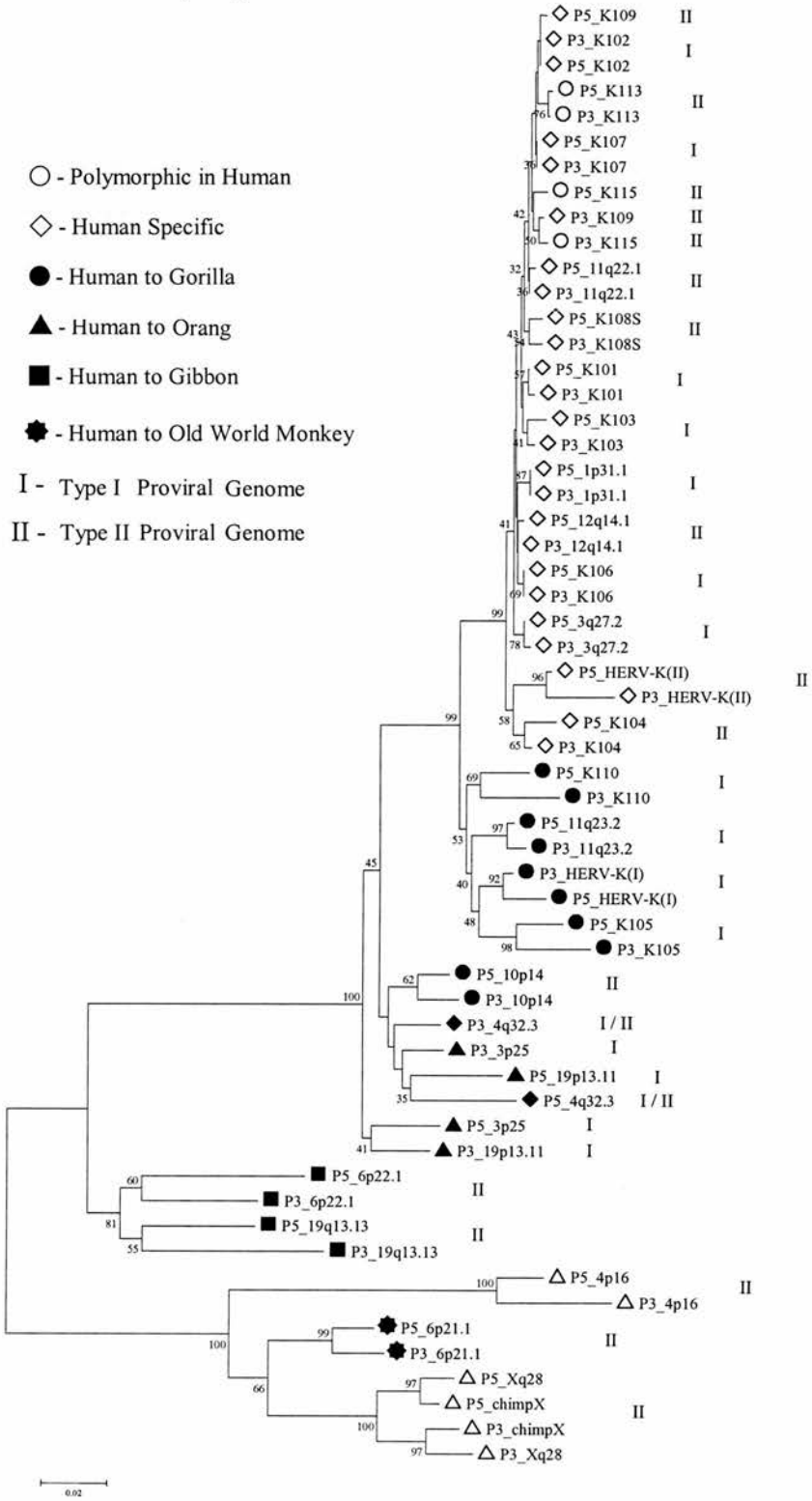


Figure 6.2 Phylogeny of the combined ORFs belonging to HERV-K(HML-2) Proviruses, showing the Topology of the Type I and Type II Proviral genomes. The neighbour-joining tree is based on the Kimura-2-parameter distance estimate.

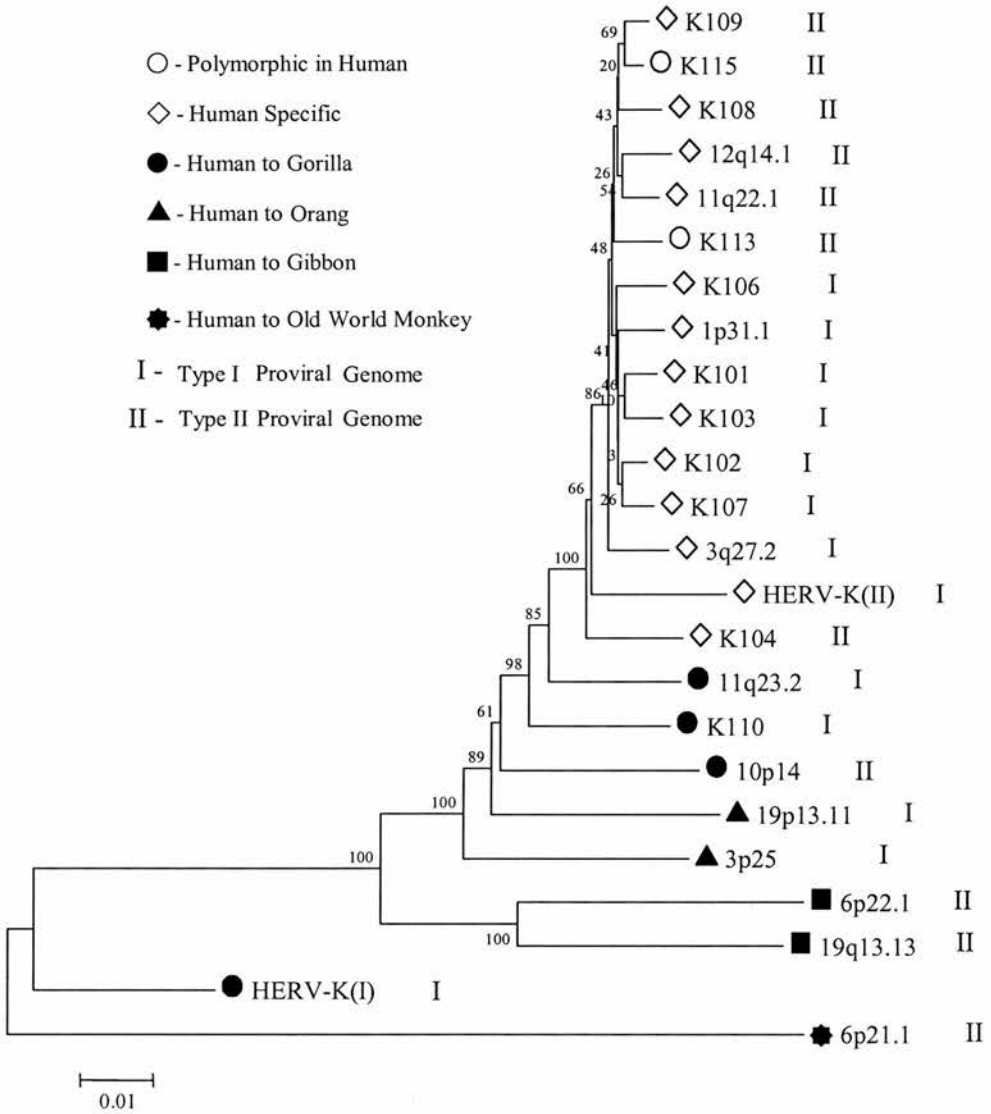


Figure 6.3 Phylogeny of the *gag* regions belonging to HERV-K(HML-2) Proviruses, showing the Topology of the Type I and Type II Proviral genomes. The neighbour-joining tree is based on the Kimura-2-parameter distance estimate.

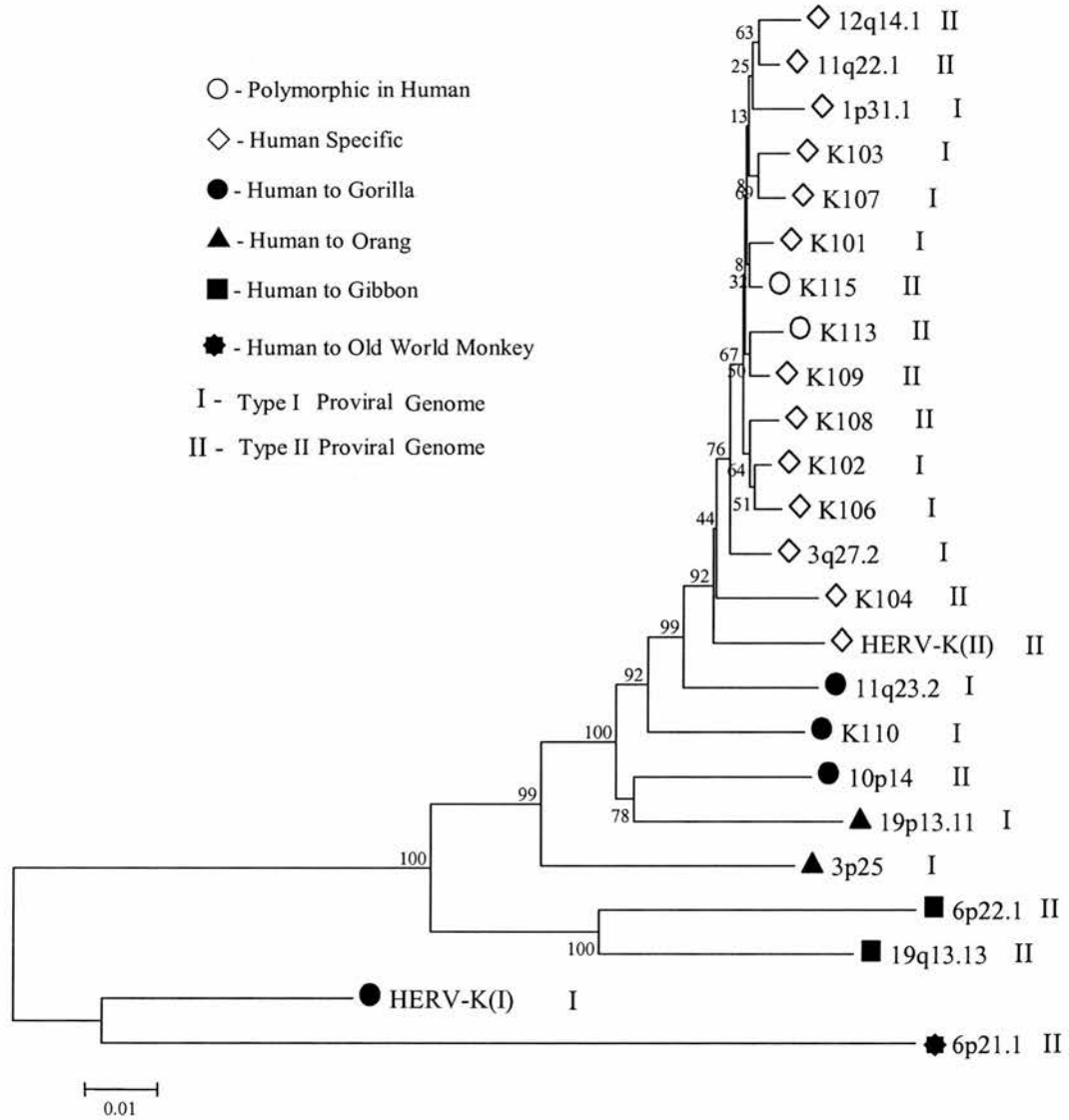


Figure 6.4 Phylogeny of the *prt* regions belonging to HERV-K(HML-2) Proviruses, showing the Topology of the Type I and Type II Proviral genomes. The neighbour-joining tree is based on the Kimura-2-parameter distance estimate.

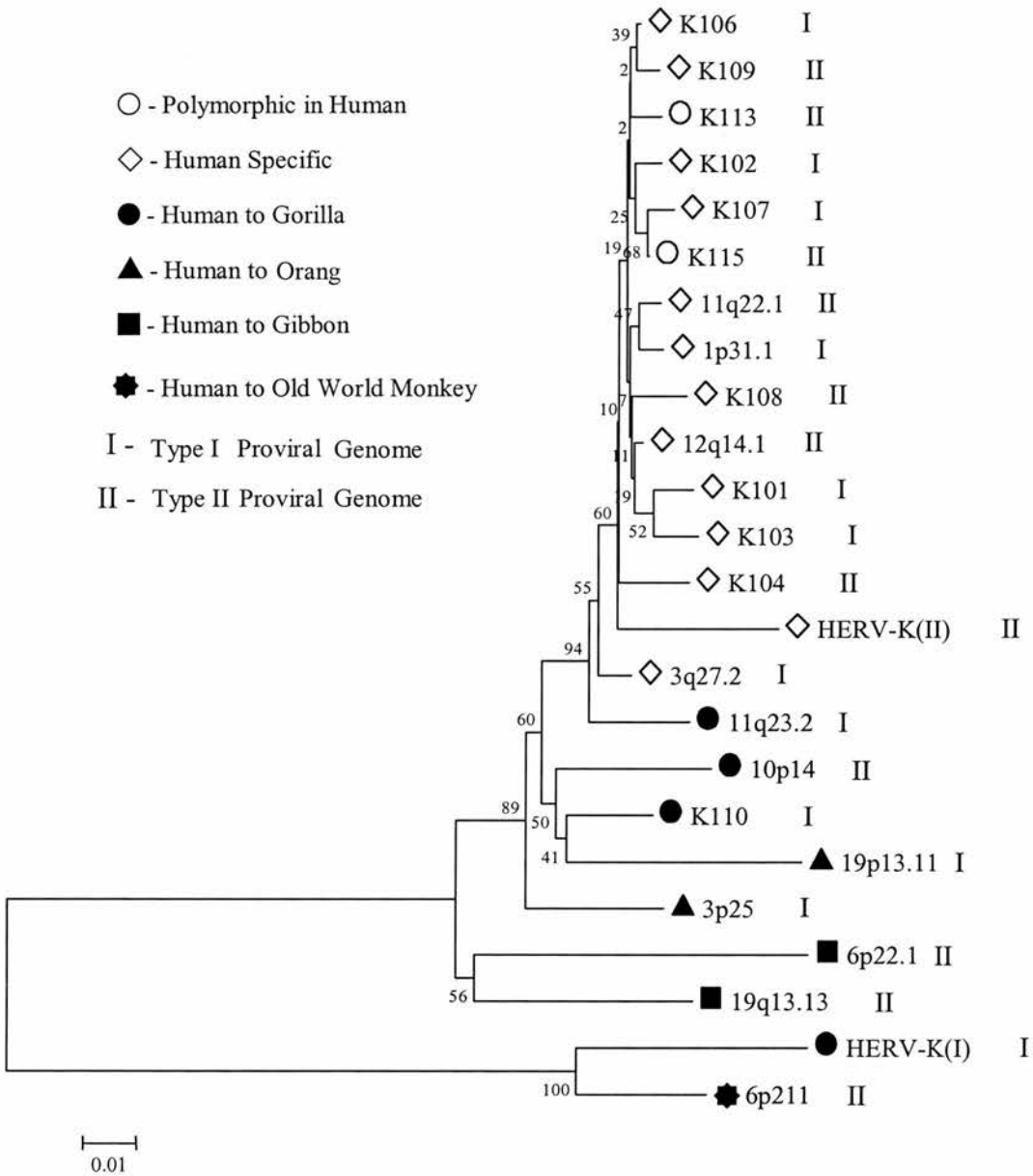


Figure 6.5 Phylogeny of the *pol* regions belonging to HERV-K(HML-2) Proviruses, showing the Topology of the Type I and Type II Proviral genomes. The neighbour-joining tree is based on the Kimura-2-parameter distance estimate.

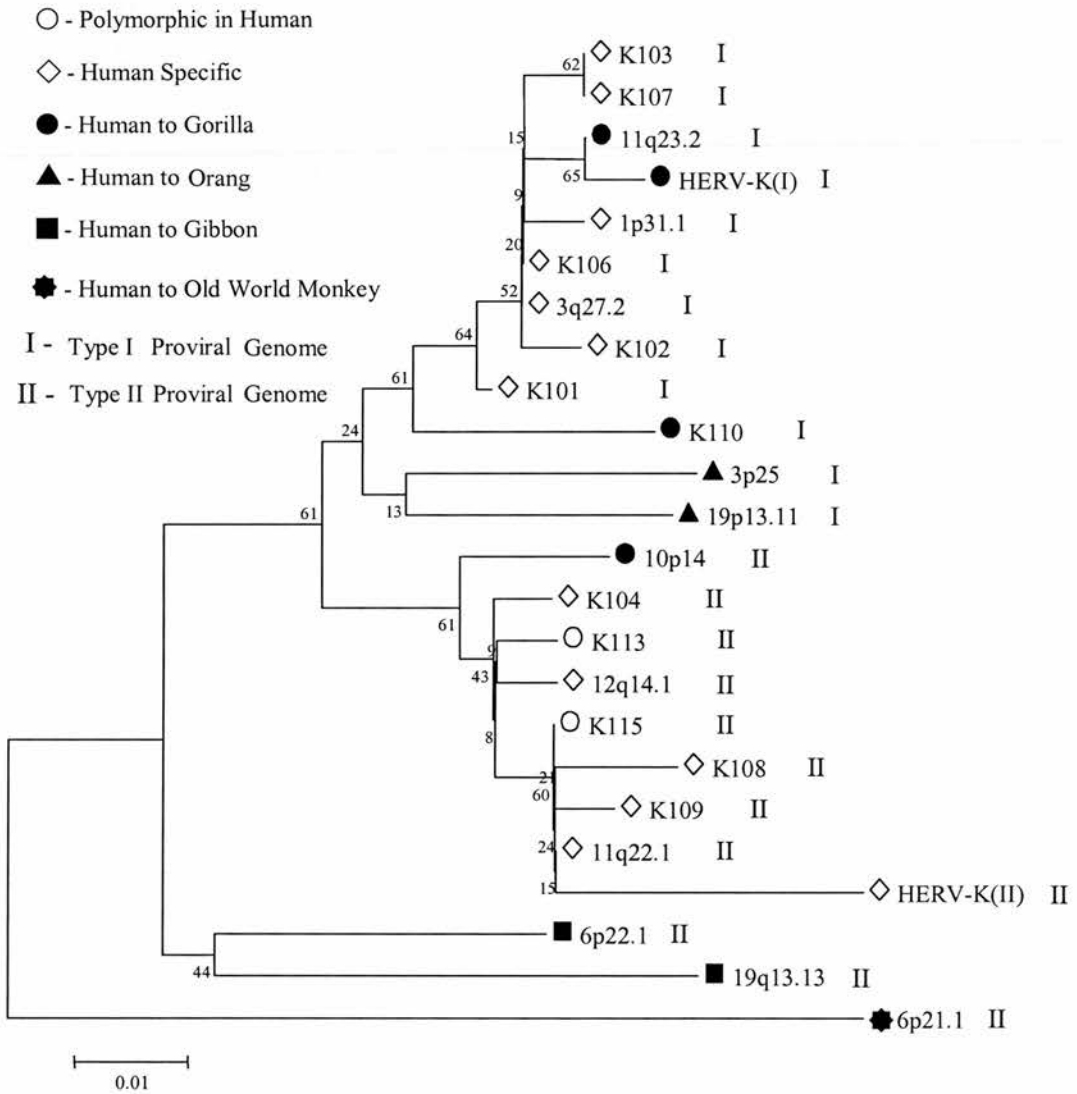
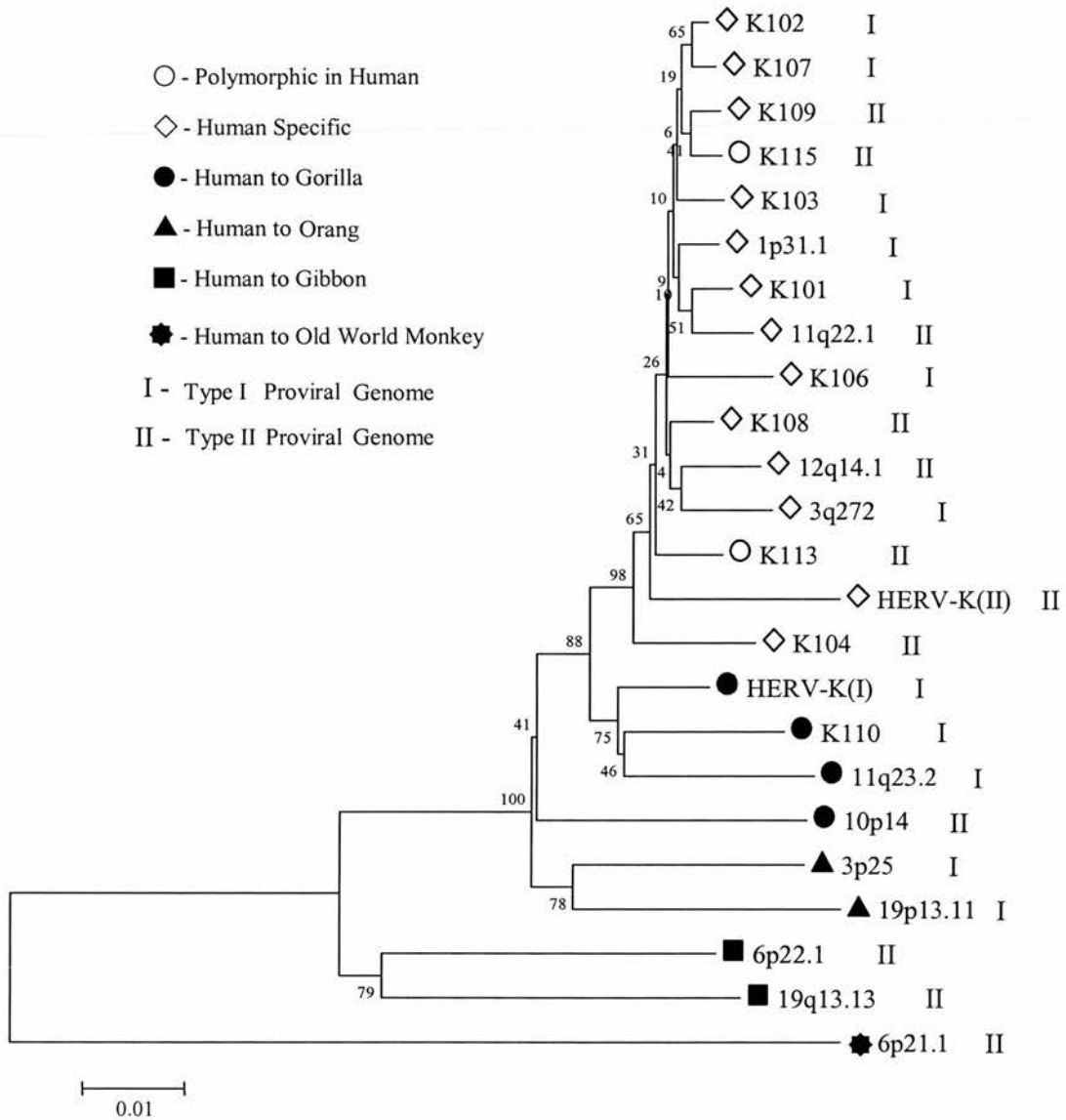


Figure 6.6 Phylogeny of the *env* regions belonging to HERV-K(HML-2) Proviruses, showing the Topology of the Type I and Type II Proviral genomes. The neighbour-joining tree is based on the Kimura-2-parameter distance estimate.



proviral lineage. The exception to this observation was the topology of the *pol* region tree (Figure 6.5). Here the proviral elements clustered into three clades, each of which represented one of the three major proviral forms, Type I, Type II or HERV-K(OLD). Such a topology is to be expected within this region as it contained the diagnostic 292 bp deletion of the Type I / Type II proviral forms. Interestingly within the three clades, elements continued to cluster consistent with relative age. If it is assumed that the polyphyletic distribution of Type I elements is due to the acquisition of the Type I deletion via sequence exchange, this result is interesting as it implies that such exchange is restricted to highly homologous sequences which are of similar relative age.

Sequence exchange between elements of similar relative age is also reflected within the LTR phylogeny. Here the proviruses, HERV-K115, HERV-K109, HERV-K 19p13.11 and HERV-K 3p25 possess LTR which do not group within the phylogenetic tree (Figure 6.1 and Chapter 5). However such exchange must have been restricted to between elements of a similar relative age as the divergent LTRs are still maintained within the same cluster. These results imply that either exchange is again restricted to highly homologous sequences or that it is taking place around the same time as proviral integration. The latter scenario is the more likely as the HERV-K115 provirus is a recent acquisition to the human genome (Chapter 4). Furthermore, analysis of orthologous loci of the divergent LTRs in non-human primates indicates that sequence exchange also occurred quickly following integration (Chapter 5).

Reconstruction of the sequence relationships of the HERV-K(HML-2) proviruses suggests that extensive sequence exchange has occurred both within LTR

and internal genic regions. However, it should be considered that the genomic retroviral elements that exist today represent only a small fraction of total germ line integration events, namely those that were not detrimental to the host and that also became fixed in the genomes of the primate lineage. Therefore, when examining the sequence relationships of retroelement families via the type of analysis presented here, it must be taken into account that the results are limited to extant sequence data and do not necessarily reflect the complete evolutionary history of the family.

6.2.2 Examination of the Ratio of Synonymous to Non-synonymous changes (dS/dN) and the Reconstruction of HERV-K sequences by Maximum Parsimony

To determine the evolutionary forces that have been acting upon the HERV-K(HML-2), HERV-K(HML-3) and HERV-K(HML-4) proviruses described within Chapter 3 and to further examine their mode of expansion, the numbers of synonymous and non-synonymous substitutions were calculated by application of the Jukes – Cantor model within the SIMMONIC sequence analysis package. The proviral sequences utilised are listed within Appendix B, Table B.6. Prior to analysis, the reliability of the Jukes – Cantor model was confirmed by comparison with the pairwise distance estimate and Nei – Gojobori method, the results are presented in Appendix B, Table B.7. Reconstruction of the minimum number of changes required to explain the HERV-K proviral and control retrotransposon datasets was achieved by application of the programme DNA PARS as part of the PHYLIP package using the SIMMONIC sequence analysis package as an interface.

As the majority of amino acids can be coded for by more than one codon, individual nucleotide substitutions can be classified as synonymous or non-synonymous. A synonymous substitution is one which does not alter the coding amino acid and a non-synonymous substitution changes the encoded amino acid. As the important biological functions of an organism are performed by proteins rather than nucleotide sequences, the rate of non-synonymous substitutions will vary according to natural selection with synonymous substitutions expected to remain neutral. As such, if a genic sequence is under negative or purifying selection, then the

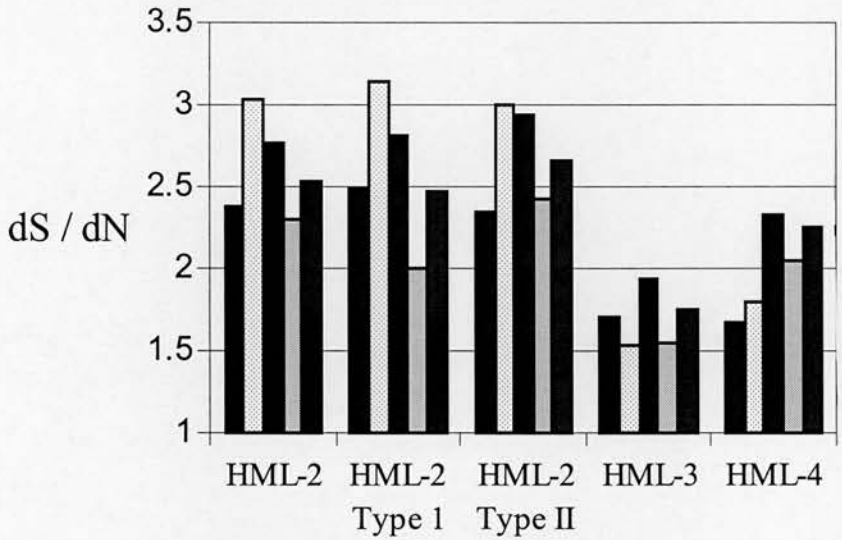
number of synonymous substitutions per synonymous site (dS) will be greater than the number of non-synonymous substitutions per non-synonymous site (dN), as the functionality of the protein sequence is being retained. Conversely, if a protein sequence is not under selection and so is evolving neutrally, the number of synonymous substitutions per synonymous site (dS) should be equal to the number of non-synonymous substitutions per non-synonymous site (dN).

Here, to evaluate the evolutionary pressures that have acted upon the three HERV-K proviral subgroups described in Chapter 3, the ratio of synonymous distances to non-synonymous distances (dS/dN) were calculated for each of the subgroups (Figure 6.7). As evolutionary forces could, in principal, differ for different proviral ORFs, *gag*, *prt*, *pol* and *env* were considered both individually and as a combined dataset. In addition, as the HERV-K(HML-2) subgroup is composed of two distinct proviral lineages, one of which is presumed to possess a functionally inactive *env* region (Type I), this subgroup was further divided into Type I and Type II proviral groups.

Overall, all HERV-K proviral groups contained more synonymous substitutions than non-synonymous substitutions, indicative of the functional preservation of their ORFs (Figure 6.7). Comparison of the combined ORFs revealed that the HERV-K(HML-3) group possessed the lowest ratio of 1.759 and HERV-K(HML-2) the highest ratio of 2.533. Subdivision of the HERV-K(HML-2) subgroup into Type I and Type II genotypes showed that for the combined ORFs, Type II elements displayed greater functional constraint than Type I proviruses, with the scores of 2.652 and 2.474 respectively.

Comparison of the individual proviral genic regions revealed that for all, the

Figure 6.7 dS/dN Ratios in the HERV-K Subfamilies HML-2, HML-3 and HML-4.

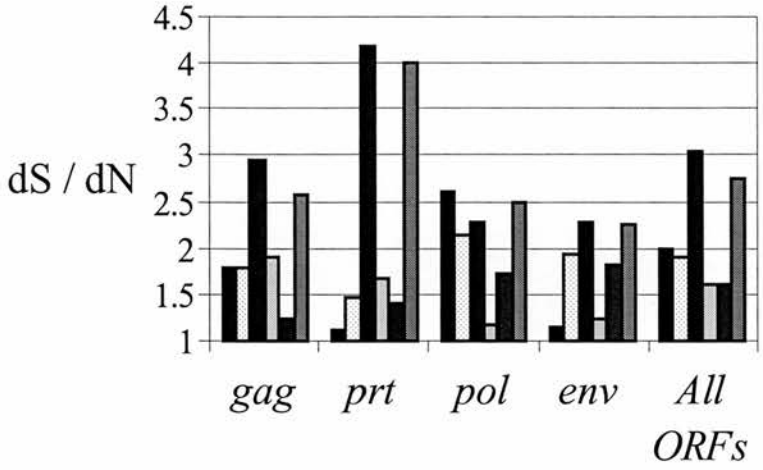


	HML-2	HML-2 Type 1	HML-2 Type II	HML-3	HML-4
■ gag	2.372	2.485	2.337	1.71	1.672
□ prt	3.019	3.14	2.99	1.54	1.801
■ pol	2.766	2.806	2.927	1.933	2.329
□ env	2.298	1.996	2.419	1.545	2.049
■ All ORFs	2.533	2.474	2.652	1.759	2.25

HERV-K(HML-2) subgroup constantly had a higher ratio of dS/dN than the HERV-K(HML-3) and HERV-K(HML-4) subgroups. Interestingly, all genic regions of the three HERV-K proviral groups had dS/dN ratios of greater than 1, suggestive that all regions had undergone more synonymous than non-synonymous substitutions, hence purifying selection (Figure 6.7). Within the *gag* and *prt* ORFs, HERV-K(HML-2) Type I proviral genomes possessed the highest dS/dN scores of 2.485 and 3.14 respectively. The HERV-K(HML-4) subgroup retained the lowest score for the *gag* region of 1.672 and the HERV-K(HML-3) proviruses, the lowest score for the *prt* region of 1.54. The HERV-K(HML-2) Type II proviral group showed the greatest functional preservation of the *pol* and *env* regions with dS/dN scores of 2.927 and 2.419. The HERV-K(HML-3) subgroup retained the lowest scores within these regions of 1.933 and 1.545. Interestingly, the *env* region of the HERV-K(HML-2) Type I proviral grouping had the second lowest dS/dN ratio of 1.996, perhaps highlighting the functional inactivity of this region when compared to the other genic regions of this proviral lineage.

As the HERV-K(HML-2) proviral subgroup retained the highest overall dS/dN scores within all genic regions and the subgroup is known to have been recently retrotranspositionally active; this subgroup was further divided into groups according to known proviral relative age and the ratio of synonymous distances to non-synonymous distances (dS/dN) was calculated for each of the genic regions (Figure 6.8). This allowed the examination of dS/dN ratios during the evolutionary divergence of the primate lineage. In addition to the grouping of relative age, the three proviruses HERV-K 6p22.1, HERV-K 19q13.13 and HERV-K 6p21.1 were also considered as an independent dataset as they were all members of the HERV-

Figure 6.8 dS/dN Ratios of the genic regions of HERV-K(HML-2) Proviruses which are grouped according to Relative Age.



	gag	prt	pol	env	All ORFs
■ Polymorphic (n = 2)	1.8	1.127	2.592	1.152	2.01
□ Human Specific (n = 13)	1.8	1.465	2.131	1.935	1.906
■ Gorilla (n = 4)	2.94	4.153	2.293	2.282	3.027
□ Orang-utan (n = 2)	1.9	1.686	1.177	1.249	1.634
■ Gibbon (n = 2)	1.25	1.418	1.732	1.823	1.616
■ HERV-K(OLD) (n = 3)	2.568	3.986	2.501	2.258	2.745

K(OLD) proviral variant.

The examination of the combined ORFs dS/dN ratios revealed that the four HERV-K(HML-2) proviruses which integrated within the common ancestor of human, chimpanzee and gorilla possessed the highest number of synonymous to non-synonymous substitutions with a score of 3.027. Interestingly, the three HERV-K(OLD) proviruses retained the second highest dS/dN ratio of 2.745, perhaps suggesting that elevated dS/dN ratios are a remnant of the progenitor exogenous HERV-K(HML-2) retrovirus. However, the overall dS/dN ratios for the genic regions within the proviruses which in relative age integrated between these two groups showed the lowest overall dS/dN scores of 1.616 (Gibbon) and 1.634 (Orang-utan).

These observations were further reflected when individual genic regions were considered. The HERV-K(OLD) proviruses constantly showed the second highest dS/dN ratios within all HERV-K(HML-2) proviral groups. With the exception of the *pol* region, the proviruses which were present at orthologous regions in all members of the Hominidae again showed the highest dS/dN ratios. Interestingly, the proviral group which possessed the highest number of synonymous to non-synonymous substitutions within the *pol* region was composed of the two proviruses which are insertionally polymorphic within contemporary humans. This could be suggestive of the recent requirement of a functional *pol* ORF. However, comparison of the remaining genic regions of the unfixed HERV-K113 and HERV-K115 proviruses to proviruses that were fixed within the human lineage revealed that the polymorphic proviral group retained the lowest dS/dN scores for the *gag*, *prt* and *env* regions. Furthermore the dS/dN scores of the *prt* (1.127) and *env* (1.152) regions were

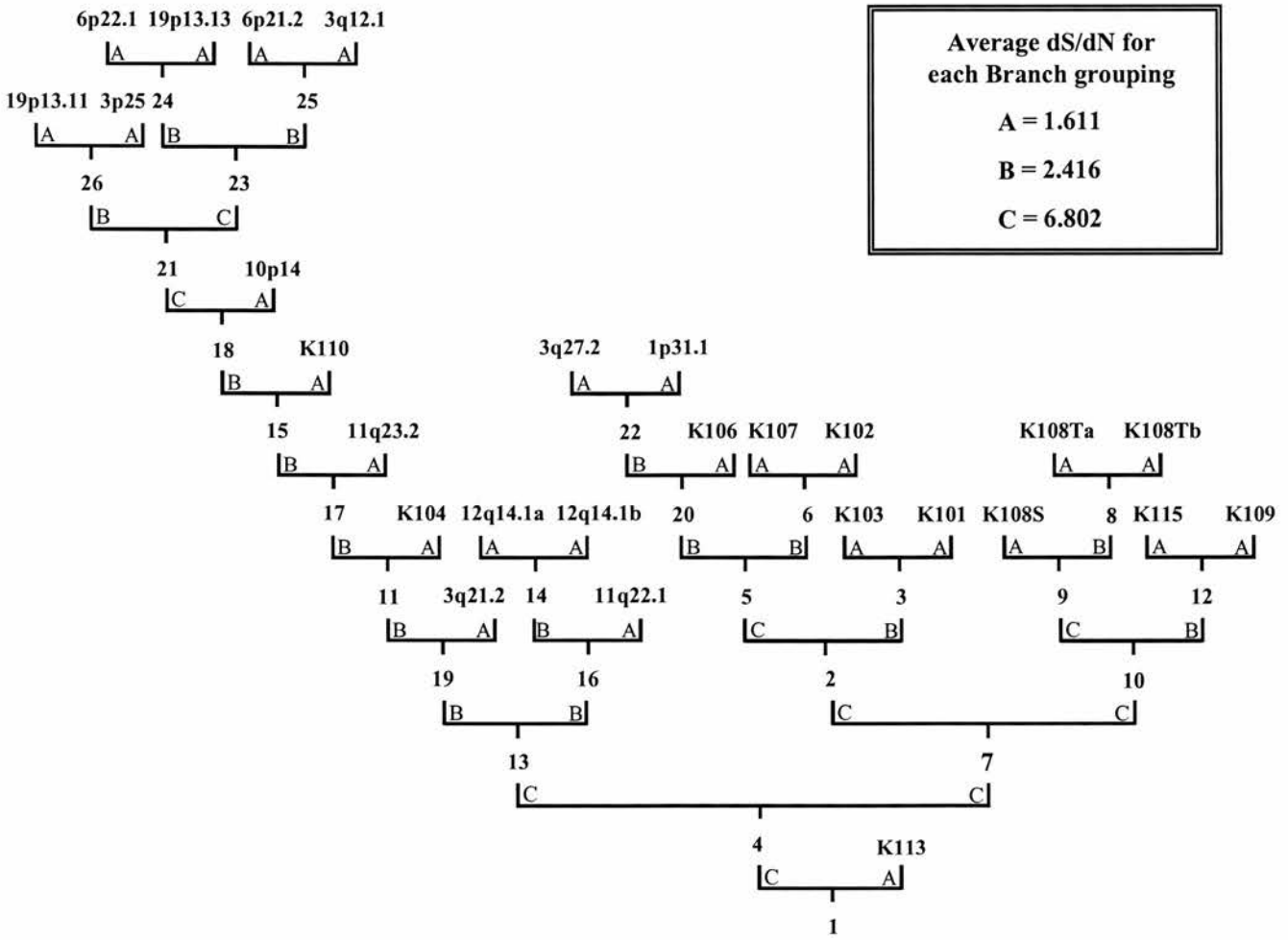
extremely close to 1, indicating that these genic regions of this group were essentially evolving neutrally and so not under purifying selection.

As HERV-K(HML-2) proviruses of greater relative age displayed a higher number of synonymous than non-synonymous substitutions when compared to proviruses of more recent integration (Figure 6.8), the elevated dS/dN ratios observed within the three HERV-K proviral subgroups (Figure 6.7) could be a remnant of purifying selection acting upon ancestral proviral sequences. Subsequently, high dS/dN ratios within proviruses of more recent relative age could be a signature of past ORF maintenance as opposed to the proviruses themselves having been subject to purifying selection.

To clarify if dS/dN was higher in the past, the phylogeny of the HERV-K(HML-2), HERV-K(HML-3) and HERV-K(HML-4) proviral ORFs were reconstructed using maximum parsimony and the ratio of dS/dN was then calculated between all phylogenetically adjacent sequences. This included between extant sequences and those reconstructed by maximum parsimony. Ratios positioned at terminal branches (extant sequences) were assigned to group 'A' and the ratios present at internal branches (reconstructed sequences) were assigned to groups 'B' or 'C' (Figure 6.9). Ratios contained within group 'C' represented those furthest removed from the extant sequences. The mean ratio of each of the groupings was then calculated to provide an indication of the level of purifying selection acting upon each of the branches of the tree.

To begin to examine the mode in which the three HERV-K proviral subgroups have expanded, control datasets consisting of both autonomous and nonautonomous retrotransposons were also reconstructed by maximum parsimony

Figure 6.9 Topology of the Maximum Parsimony Tree of HERV-K(HML-2) Proviral ORFs. Reconstructed sequences are represented by numbers. Terminal Branches are annotated A, Internal Branches B and C.

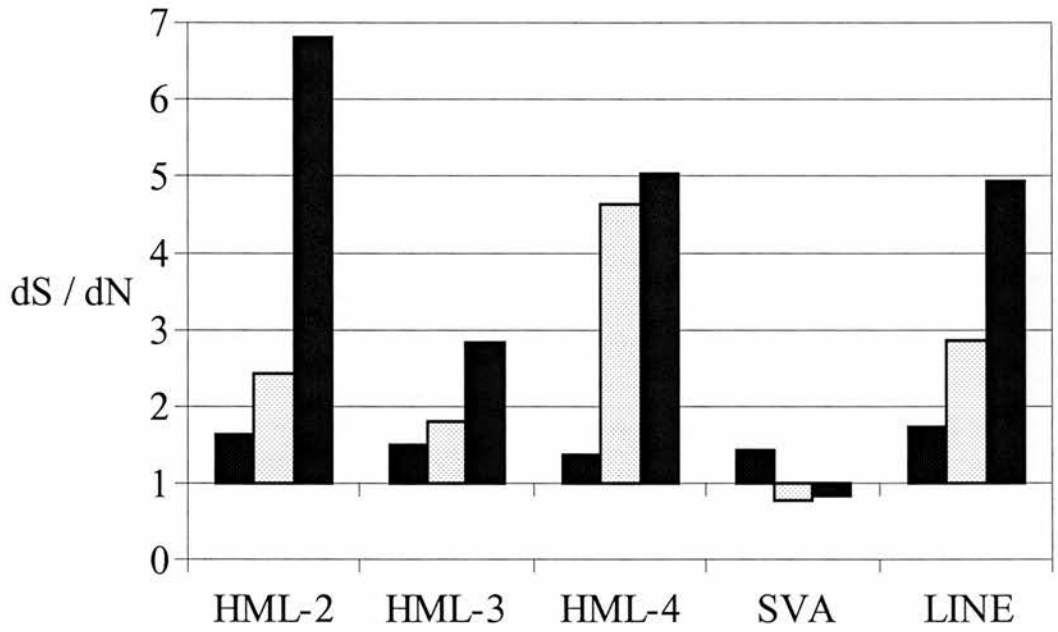


and the mean dS/dN ratios calculated for each of the branches. The first of the control datasets consisted of the two ORFs of the LINE retrotransposon family which is acknowledged to replicate in *cis*. Analysis of the dS/dN ratio of these regions within 192 intact extant LINE elements gave a score of 1.865. The second of the control datasets consisted of 99 bp (33 codons) of the (HERV-K(HML-2) *env* region which is retained within the SVA retrotransposon family. Examination of the dS/dN ratio of this region within 92 intact extant SVA elements provided a score of 1.159, confirming that that this retro-element family replicate in *trans*.

Overall, the mean dS/dN ratios for each of the phylogenetic reconstructions of the three HERV proviral subgroups indicated that the dS/dN ratios were highest within the internal branches and lowest in the terminal branches. This suggests that functional constraint of the proviral ORFs observed within the terminal branches (extant sequences) is a remnant of former purifying selection (Figure 6.10). Of the three proviral datasets, the HERV-K(HML-2) group displayed the highest dS/dN score of 6.806 within the internal branch 'C'. Interestingly, the HERV-K(HML-4) group retained the highest dS/dN ratio within the internal branch 'B' of 4.643 although this group possessed the lowest mean dS/dN ratio of 1.372 within the terminal branch. The HERV-K(HML-3) group displayed an increase in the number of synonymous to non-synonymous substitutions within the internal branches of the phylogeny; however the mean scores of 1.776 and 2.825 were low when compared to the equivalent branches of the HERV-K(HML-2) and HERV-K(HML-4) groups.

The internal branches of the non-autonomous SVA elements showed that the number of synonymous substitutions was less than the number of non-synonymous substitutions with a mean score of 0.753 for the internal 'B' branch and 0.808 for the

Figure 6.10 dS/dN Ratios of the Terminal and Internal Branches of Maximum Parsimony Trees.



	HML-2	HML-3	HML-4	SVA	LINE
■ Terminal Branch A	1.611	1.502	1.372	1.41	1.719
□ Internal Branch B	2.416	1.776	4.643	0.753	2.859
■ Internal Branch C	6.806	2.825	5.029	0.808	4.935

internal 'C' branch. As the region analysed was within the past acquired from an HERV-K(HML-2) element, this result is significant as it emphasises that the functionality of the *env* ORF has not been retained within this retrotransposon family although the terminal branches have a mean dS/dN score of 1.41 (Figure 6.10).

The autonomous LINE elements displayed a similar elevation of dS/dN ratios to the HERV-K proviral subgroups within the internal branches, with scores of 2.859 and 4.935 respectively. Furthermore, the LINE terminal branches (extant sequences) had a mean of 1.719 which was higher than the terminal ratio scores for the three proviral subgroups. As LINE elements replicate in *cis*, it is possible that these dS/dN scores reflect the rapid replicative ability of this family.

Analysis of the evolutionary pressures that have been acting upon the HERV-K(HML-2), HERV-K(HML-2) and HERV-K(HML-4) proviral subgroups revealed that all genic regions appeared to be functionally constrained (Figure 6.7). Subdivision of the HERV-K(HML-2) proviruses into groups of relative age further showed that, with the exception of the *pol* region, this constraint could be a remnant of past purifying selection (Figure 6.8). Reconstruction of the proviral phylogenies and analysis of dS/dN ratios within internal and terminal branches confirmed this view (Figures 6.9 and 6.10). Finally, comparison of the scores of the internal and terminal proviral branches to datasets constructed from both autonomous and non-autonomous retrotransposon families revealed that the levels of selection acting upon the reconstructed proviral phylogenies were most similar to those of a retrotransposon family which replicates in *cis* (Figure 6.10).

6.3 Discussion

It has previously been suggested that the slower rate of non-synonymous substitutions compared to synonymous substitutions across the entire ORFs of the HERV-K family is indicative of purifying selection (Zsiros et al., 1998; Zsiros et al., 1999; Costas, 2001; Belshaw et al., 2004). Furthermore, the conclusions drawn from more recent analysis indicate that the HERV-K(HML-2) subfamily has proliferated via germ-line reinfection as opposed to retrotransposition in *cis* or in *trans* (Belshaw et al., 2004). Within this recent publication, proliferation in *trans* was excluded as stop codons appeared to be rarely inherited and all internal branches of the reconstructed ORF phylogeny showed dS/dN ratio of > 1. As the *env* region displayed dS/dN scores of > 1 on both terminal and internal branches and the inheritance of stop codons within this region was rare, expansion in *cis* was also ruled out. Furthermore, the authors proposed that within a reconstructed phylogeny, if the family was following a 'master element' model of replication (*cis*), the internal nodes should represent the same master element integration at different times and that the dS/dN ratios on the internal branches would not be significantly different from 1 (Belshaw et al., 2004).

Within this study the proliferation of the HERV-K(HML-2), HERV-K(HML-3) and HERV-K(HML-4) proviral lineages was considered by analysis of dS/dN across their ORFs, reconstruction of their phylogeny with analysis of dS/dN between adjacent sequences and comparison to retrotransposon families which expand in *trans* (Kim et al., 1999; Ostertag et al., 2003) and *cis* (Wei et al., 2001). The results obtained here for the HERV-K families are congruent with those previously reported

and show the functional constraint of ORFs. However, within this study, analysis of the levels selection acting upon LINE elements demonstrated that expansion in *cis* cannot be ruled out as a mechanism of HERV-K(HML-2) element expansion.

First of all, it was observed that the LINE element ORFs were highly conserved and that strong purifying selection had been acting upon these regions as a dS/dN ratio of 1.865 was obtained. This result is surprising as the vast majority of full length LINE elements are 'dead end' or replication incompetent. However, such a phenomenon has also been observed within members of the L1PA5 to L1PA1 families (Boissinot and Furano, 2001) which have expanded over the last 25 Mya (Boissinot et al., 2000). This trend is likely to be attributed to the selective pressure acting upon the progenitor 'master' elements (Hardies et al., 1986; Brouha et al., 2003) whereby their progeny inherit the remnants of this selection. Second, reconstruction of the LINE element phylogeny revealed higher levels of purifying selection acting upon internal than terminal nodes. This result is in direct opposition to the expectation of master element expansion described in Belshaw et al., (2004). However, as the extant terminal nodes are descended from elements which were under selective pressure to maintain protein expression (Hardies et al., 1986), these elements could have inherited the signatures of purifying selection. Furthermore, a large dataset of full length LINE elements (192) was utilised within this study which were also divergent at the 1 % level, so it is possible that the internal nodes within the 'C' grouping represent more than one 'master' element integration.

Analysis of the levels of selection acting upon the nonautomomous (*trans*) SVA retrotransposon family was also informative. The region analysed was acquired from the HERV-K(HML-2) proviral lineage (Ono et al., 1987) and conveniently

corresponds to the last 99 bp of the *env* region. As would be expected from a retrotransposon family that replicates in *trans*, which therefore does not require the *env* region for retrotransposition, the dS/dN ratio for the 92 SVA elements was close to 1 (1.159). Interestingly, reconstruction of the internal nodes of this family revealed that the rate of non-synonymous substitutions was slightly higher than the rate of synonymous substitutions, indicating that this region may have been undergoing positive selection in the past. It is possible that this is due to the young age of this family, < 15 Mya (Kim et al., 1999), whereby the family is under selective pressure to evade repression by the host genome. Such a scenario has been proposed for the (217 bp) coiled coil domain within ORF1 of LINE elements belonging to the L1PA5 to L1PA3B families (Boissinot and Furano, 2001).

An interesting result obtained both within this study and in previous publications is the apparent maintenance of the HERV-K(HML-2) Type I *env* ORF (Costas, 2001; Belshaw et al., 2004). Here, Type I elements have a dS/dN ratio of 1.996, which is comparable to the scores of 2.416, 1.545 and 2.049 obtained for the *env* regions of HERV-K(HML-2) Type II, HERV-K(HML-3) and HERV-K(HML-4) proviral lineages. As the *env* region of the HERV-K(HML-2) Type I proviral lineage is presumed to be non-functional as a result of an altered splicing pattern, this result is surprising. Analysis of inherited stop codons within this *env* region has shown that they are rare, indicating that this proviral lineage has not been transmitted in *cis* or in *trans* (Belshaw et al., 2004).

Intriguingly, the distribution of the Type I and Type II proviral forms was not monophyletic within trees constructed from the LTR, *gag, prt, env* and combined ORFs (Macfarlane and Simmonds, 2004) as would be expected from a clonal

expansion model. In addition, LTR subtypes HS-a and HS-b grouped independently from the proviral variants within the LTR tree (Macfarlane and Simmonds, 2004). Conversely, Type I and Type II proviral variants were polyphyletic within the tree constructed from the *pol* ORF which encompassed the diagnostic 292 bp region and clustered according to relative age suggesting a common origin of this deletion. Furthermore, all observed tree topologies were consistent when gaps within the sequence data were handled as both a complete deletion and pairwise deletion (data not shown). This, in concert with observations in Chapters 3, 4 and 5, suggests that HERV-K(HML-2) sequences have been subject to a high degree of sequence exchange between highly homologous sequences.

Subdivision of the HERV-K(HML-2) proviruses into groups of relative age and calculation of the dS/dN ratio within each of the ORFs indicated that older proviral variants contained the highest dS/dN ratios of all proviral groupings. This implies that proviral sequences of greater relative age were subject to the strongest levels of purifying selection. Interestingly, with the exclusion of the *pol* region, proviruses of more recent relative age showed the lowest dS/dN ratios for the *gag, prt* and *env* regions (close to 1), indicating that these regions have not recently been subject to purifying selection. These observations insinuate that dS/dN ratios higher than 1, within these three genic regions, could be a remnant of past selection acting upon the progenitor (exogenous retrovirus) sequences. Consequently, the apparent functional constraint of the HERV *gag, prt* or *env* regions might not be a signature of recent purifying selection.

However, the dS/dN ratios obtained for the group of HERV-K(HML-2) proviruses which are present within all members of the Hominidae family, were

extremely low. As the representatives of this group, HERV-K 3p25 and HERV-K 19p13.11a, were previously observed to have undergone sequence exchange (Chapter 5) and they share a 1937 bp deletion of the *pol* region (Chapter 3), it is possible that they have undergone concerted evolution following integration which would invalidate the capability of calculating synonymous and non-synonymous substitution distances. The concerted evolution of retrotranposon families has been observed within LINE pseudogenes (Hardies et al., 1986), HERV-K proviruses (Dangel et al., 1995; Johnson and Coffin, 1999) and HERV-H proviruses (Mager and Freeman, 1995).

Furthermore, it is possible that the elevated dS/dN ratios within the older proviral groups could be an artefact of high sequence divergence, whereby the sequences have reached substitution saturation, which leads to a loss of phylogenetic signal so the underlying evolutionary processes which produced them cannot be determined (Seifarth et al., 1995). However, the reliability of the dS/dN ratios obtained for the different proviral groups was confirmed by application of different models of estimating sequence distances which considered different substitution patterns (Appendix B, Table B.7). In addition, the dS/dN ratios obtained for the SVA elements were close to 1 as would be expected from a retrotransposon family which replicates in *trans*.

Reconstruction of the phylogeny of the HERV-K(HML-2), HERV-K(HML-3) and HERV-K(HML-4) using maximum parsimony further confirmed that dS/dN ratios were higher in the past. The integrity of this technique was confirmed as the topologies of the proviral subgroups constructed by neighbour-joining (Chapter 5) were congruent with those reproduced by maximum parsimony (data not shown).

However, the application of further statistical tests, for example likelihood, is required in order to assess the true significance of dS/dN ratios on different branches of the reconstructed phylogenetic tree.

There are several observations that support the hypothesis that the HERV-K(HML-2) subgroup proliferated within germ-line cells via reinfection over the last 30 Mya. First, they have retained the ability to encode functional retroviral protein (Towler et al., 1998; Berkhout et al., 1999; de Parseval et al., 2003; Mayer et al., 2004). Second, they are associated with retrovirus-like particles (Peakall and Smouse, 2001; Seifarth et al., 1995; Simpson et al., 1996; Bieda et al., 2001). Finally, their ORFs appear to be maintained and includes the *env* region, which is presumed to be only required for movement between cells (Zsiros et al., 1998; Zsiros et al., 1999; Costas, 2001; Belshaw et al., 2004). However, within this study, examination of the evolutionary forces that have acted upon this family over the last ~ 30 Mya suggests that elevated dS/dN ratios could be a remnant of the initial infection by a pool of exogenous retrovirus. In addition, expansion of this subfamily in *cis* cannot be excluded as the forces acting upon the reconstructed phylogeny are very similar to those of the autonomous LINE retrotransposon family which contain members which are replication incompetent. Further examination of the evolutionary pressures that have been acting upon other HERV families, such as the HERV-W family which is reported to have increased in number within the human genome via retroviral transposition and LINE mediated retrotransposition (Costas, 2002; Pavlicek et al., 2002), will resolve the mode in which the HERV-K family has expanded throughout the evolutionary divergence of the primates.

CHAPTER 7

FINAL DISCUSSION

7 Final Discussion

Endogenous retroviruses (ERVs) are the remnants of ancient germ cell infection by exogenous retroviruses and occupy up to 8 % of the human genome. Following initial infection approximately 28 Million years ago, members of the HERV-K family have continued to amplify and recombine within the genomes of the primate lineage. The role of HERV-K in primate evolution is yet to be fully determined, however it has been proposed that they may have contributed by conferring resistance to retroviral infection, serving as mediation points for chromosomal rearrangements and acting as regulators of host gene expression. Furthermore, the mode in which this family have proliferated within the genomes of the primate lineage is poorly understood and the longevity of this expansion unspecified.

The near completion of the human genome sequencing project provides a unique starting point for addressing these issues as the structure and cytogenetic location of HERV-K elements can be easily determined. Moreover, as the retrotransposition and insertion of a HERV sequence within the germ line represents a unique event in primate genome evolution, the relative age of each element can be phylogenetically determined by amplification in extant primate lineages. This data can be combined to examine the retrotranspositional history of the HERV-K family.

In this study a comprehensive catalogue of intact and near intact HERV-K(HML-2), HERV-K(HML-3) and HERV-K(HML-4) proviral sequences that are present within the human genome was determined. In addition, the genomic location and total number of previously reported HERV-K(HML-2) solitary LTRs was also

ascertained. As well as highlighting numerous inconsistencies within the literature, six novel HERV-K proviruses and a diagnostic region within the *gag* ORF of the HERV-K(HML-2) proviruses were identified.

The validity of LTR divergence of individual elements in serving as a molecular clock was considered by comparing the LTR estimated age of integration to the relative age as determined by their presence or absence in non-human primates. Of the 27 HERV-K(HML-2) proviruses compared, three possessed LTRs whose divergence was significantly less than would be expected according to their relative age. Thus LTR divergence might not always serve as an accurate indicator of time passed since integration.

As the relative age of each of the elements was determined, the retrotranspositional history of the HERV-K(HML-2) proviral lineages could be examined. The results indicated that the HERV-K(OLD) variant ceased to amplify following the evolutionary split of gorillas from the human lineage with the shorter *gag* variant arising following the divergence of the Cercopithecoidea and Hominoidea super families. The HERV-K(HML-2) Type I genotype, which is presumed to be non-functional as a result of a 292 bp deletion within the *pol-env* boundary, was determined to have arisen following the evolutionary split of gibbons from the human lineage with amplification continuing following the evolutionary divergence of human and chimpanzee. Interestingly, of the 19 elements which belonged to the HERV-K(HML-2) Type II genotype, two were insertionally polymorphic within humans, suggesting the very recent activity of this lineage.

ERVs serve as ideal phylogenetic markers for examining primate evolution and speciation for a number of reasons. Firstly, the acquisition of an ERV within the

germ line represents a unique event in genome evolution and is transmitted as a Mendelian trait in succeeding generations. Secondly, HERVs are homoplasy free traits for which there are no known mechanisms of complete removal without resulting in a telltale deletion of host chromosomal DNA or production of a solitary LTR. Accordingly, the directionality of the insertion and the formation of the solitary LTR can unambiguously be assigned to a specific lineage. Finally, the ancestral state of the HERV is ultimately its absence and is represented by a pre-integration site sequence, this can be used to root trees of inter and intra population relationships.

As a starting point for examining the utility of HERVs in serving as phylogenetic markers, each of the catalogued HERV-K elements was screened within the human genome databases for variability. The results showed that 7 loci were variable, all of which were human specific integrations and belonged to the HERV-K(HML-2) subgroup. Two of the loci were solitary LTRs which were polymorphic for insertion. The first, HERV-K 6p21.32, is reported to have arisen through the duplication of the MHC complex and so does not represent a recent retrotranposition event. The second, HERV-K 9q12 was located within a highly repetitive chromosomal location so it was impossible to determine if it was genuinely insertionally polymorphic. Of the remaining five loci, two were variable for the insertion of a provirus, two for the alternation of a solitary LTR and a provirus and the fifth was variable for a tandemly repeated provirus or single provirus.

Polymerase chain reaction based assays were developed for each of the five variable loci and also for a further two proviral loci which appeared to be monomorphic within the human genome databases. Following this, 109 human DNA samples from Africa, Europe, Asia and Southeast Asia were screened to determine

global allelic variation and statistical analysis was conducted to examine the inter and intra population relationships. The results indicated that 90.12 % to 99.37 % of genetic variation was within a population suggesting that contemporary humans are a very homogenous species. Furthermore, the two most geographically distant population groups, the Africans and Papua-New-Guineans, were the most closely related. Further analysis of heterozygosity levels indicated that the Papua-New-Guineans had remained isolated for a long period. In concert, these results suggest an 'Out of Africa' model of contemporary human dispersal. However, the African population retained the highest heterozygosity level, which is not in keeping with this model. Moreover, such a result is suggestive of either Africa retaining a large long-term effective population size or extensive gene flow back into Africa, both of which are consistent with the 'Multiregional model' of human dispersal. The results obtained within this study showed that variable HERV-K loci do serve as highly sensitive markers for examining the evolution and dispersal of population groups. Furthermore they emphasised that contemporary human populations arose in Africa but they were subject to a complex and potentially long-term process of interbreeding and population movement.

Investigation of the biological contribution of HERV sequences in serving as nucleation points for chromosomal rearrangement demonstrated that such events have been extremely rare during primate evolution and that their frequency may have been overestimated in the past. It has previously been suggested that proviruses that retain variable target site duplications and possess LTRs which do not cluster within a phylogenetic tree are the end products of inter-element recombination events, which will have resulted in the translocation of chromosomal DNA (Hughes and

Coffin, 2001). However, the results obtained here show that target site duplications can vary as a result of nucleotide substitution and that the LTRs of an individual element can be highly divergent although their target site duplications remain identical. Conversely, an element can possess LTRs which cluster together on a phylogenetic tree but retain disparate target site duplications. Such outcomes are attributed within this study to sequence exchange between highly homologous, but directly unrelated, LTRs and sequence homogenisation within an individual element. Furthermore, specific investigation of human specific HERV-K(HML-2) elements which had variable direct repeats revealed that unequal crossover and deletion of a few nucleotides will also lead to disparate target site duplications.

Further analysis of the HERV-K(HML-2) subgroup revealed that their coding regions had also been subject to extensive sequence exchange throughout their expansion. Interestingly, the Type I elements were not monophyletic within trees constructed from different regions as would be expected from a clonal expansion model. In addition, the LTR subtypes Hs-a and Hs-b grouped independently from the proviral variants within a tree constructed using LTRs (Macfarlane and Simmonds, 2004). However, Type I elements did form a monophyletic group within a tree constructed using the region encompassing the diagnostic 292 bp region, which defines the Type I and Type II proviral variants. This implies a common origin for the Type I genotype 292 bp deletion. Interestingly, within all trees elements of a similar age clustered together irregardless of their HERV-K(HML-2) proviral form, this further suggests sequence exchange between highly homologous sequences.

HERV sequences which are present within the human genome are recognised to have been initially acquired via ancient retroviral infection of germ line cells.

Following the original insertion of a provirus, intracellular retrotransposition in *cis* or *trans* and reinfection have been proposed as the mechanisms by which particular families have proliferated. The mode of HERV-K proliferation was considered within this study by analysis of the synonymous (dS) and non-synonymous (dN) changes across their ORFs, reconstruction of their phylogeny with analysis of dS/dN between adjacent reconstructed and extant sequences and comparison to the retrotransposon families LINE and SVA. The levels of selection acting upon LINE elements demonstrated that expansion in *cis* cannot be excluded as a mechanism of HERV-K proliferation. Furthermore, subdivision of HERV-K(HML-2) proviruses into groups of their determined relative age showed that proviruses of greater relative age were, in the past, under greater purifying selection than those of more recent acquisition. This observation is noteworthy as it indicates that constraints on sequence variation have reduced over time, suggesting a decline in the likelihood of HERV functionality. Conversely, analysis of the ORFs of the 52 HERV-K proviruses detected within this study indicated that three, all of which were unique to humans, may have retained the ability to enocode retroviral proteins.

Whilst the role of HERVs in primate evolution is yet to be fully understood, the comprehensive catalogue obtained within this study, the identification of novel proviral sequences and further elucidation of recombinant events provide the foundations for future functional and phylogenetic investigations of human and primate evolution and speciation.

REFERENCES

References

Adcock,G.J., Dennis,E.S., Easteal,S., Huttley,G.A., Jermiin,L.S., Peacock,W.J., and Thorne,A. (2001). Mitochondrial DNA sequences in ancient Australians: Implications for modern human origins. Proc. Natl. Acad. Sci. U. S. A *98*, 537-542.

Agusti,J., Sanz,d.S., and Garces,M. (2003). Explaining the end of the hominoid experiment in Europe. J. Hum. Evol. *45*, 145-153.

Aiello,L.C. and Collard,M. (2001). Palaeoanthropology Our newest oldest ancestor? Nature *410*, 526-527.

Akopov,S.B., Nikolaev,L.G., Khil,P.P., Lebedev,Y.B., and Sverdlov,E.D. (1998). Long terminal repeats of human endogenous retrovirus K family (HERV-K) specifically bind host cell nuclear proteins. FEBS Lett. *421*, 229-233.

Al Sumidaie,A.M., Leinster,S.J., Hart,C.A., Green,C.D., and McCarthy,K. (1988). Particles with properties of retroviruses in monocytes from patients with breast cancer. Lancet *1*, 5-9.

Alemseged,Z., Coppens,Y., and Geraads,D. (2002). Hominid cranium from Omo: Description and taxonomy of Omo-323-1976-896. Am. J. Phys. Anthropol. *117*, 103-112.

Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W., and Lipman,D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. *25*, 3389-3402.

Andersson,G., Svensson,A.C., Setterblad,N., and Rask,L. (1998). Retroelements in the human MHC class II region. Trends Genet. *14*, 109-114.

Andersson,M.L., Lindeskog,M., Medstrand,P., Westley,B., May,F., and Blomberg,J. (1999). Diversity of human endogenous retrovirus class II-like sequences J. Gen. Virol. *80 (Pt 1)*, 255-260.

Andersson,M.L., Medstrand,P., Yin,H., and Blomberg,J. (1996). Differential expression of human endogenous retroviral sequences similar to mouse mammary tumor virus in normal peripheral blood mononuclear cells. AIDS Res. Hum. Retroviruses *12*, 833-840.

Armbruester,V., Sauter,M., Krautkraemer,E., Meese,E., Kleiman,A., Best,B., Roemer,K., and Mueller-Lantzsch,N. (2002). A novel gene from the human endogenous retrovirus K expressed in transformed cells. Clin. Cancer Res. *8*, 1800-1807.

Arnason,U., Gullberg,A., Janke,A., and Xu,X. (1996). Pattern and timing of evolutionary divergences among hominoids based on analyses of complete mtDNAs. J. Mol. Evol. *43*, 650-661.

Asfaw,B., Gilbert,W.H., Beyene,Y., Hart,W.K., Renne,P.R., WoldeGabriel,G., Vrba,E.S., and White,T.D. (2002). Remains of Homo erectus from Bouri, Middle Awash, Ethiopia. Nature *416*, 317-320.

Asfaw,B., White,T., Lovejoy,O., Latimer,B., Simpson,S., and Suwa,G. (1999). Australopithecus garhi: a new species of early hominid from Ethiopia. Science *284*, 629-635.

Awadalla,P., Eyre-Walker,A., and Smith,J.M. (1999). Linkage disequilibrium and recombination in hominid mitochondrial DNA. Science *286*, 2524-2525.

Babcock,M., Pavlicek,A., Spiteri,E., Kashork,C.D., Ioshikhes,I., Shaffer,L.G., Jurka,J., and Morrow,B.E. (2003). Shuffling of genes within low-copy repeats on 22q11 (LCR22) by Alu-mediated recombination events during evolution. Genome Res. *13*, 2519-2532.

Bachtrog,D. (2003). Adaptation shapes patterns of genome evolution on sexual and asexual chromosomes in Drosophila. Nat. Genet. *34*, 215-219.

Bailey,J.A., Liu,G., and Eichler,E.E. (2003). An Alu transposition model for the origin and expansion of human segmental duplications. Am. J. Hum. Genet. *73*, 823-834.

Bannert,N. and Kurth,R. (2004). Retroelements and the human genome: new perspectives on an old relation. Proc. Natl. Acad. Sci. U. S. A *101 Suppl 2:14572-9. Epub; 2004 Aug 13.*, 14572-14579.

Barbulescu,M., Turner,G., Seaman,M.I., Deinard,A.S., Kidd,K.K., and Lenz,J. (1999). Many human endogenous retrovirus K (HERV-K) proviruses are unique to humans. Curr. Biol. *9*, 861-868.

Barbulescu,M., Turner,G., Su,M., Kim,R., Jensen-Seaman,M.I., Deinard,A.S., Kidd,K.K., and Lenz,J. (2001). A HERV-K provirus in chimpanzees, bonobos and gorillas, but not humans. Curr. Biol. *11*, 779-783.

Bartsiokas,A. (2002). Hominid cranial bone structure: a histological study of Omo 1 specimens from Ethiopia using different microscopic techniques. Anat. Rec. *267*, 52-59.

Batzer,M.A. and Deininger,P.L. (2002). Alu repeats and human genomic diversity Nat. Rev. Genet. *3*, 370-379.

Batzer,M.A., Stoneking,M., Alegria-Hartman,M., Bazan,H., Kass,D.H., Shaikh,T.H., Novick,G.E., Ioannou,P.A., Scheer,W.D., Herrera,R.J., and . (1994). African origin of human-specific polymorphic Alu insertions. Proc. Natl. Acad. Sci. U. S. A *91*, 12288-12292.

Baust,C., Seifarth,W., Germaier,H., Hehlmann,R., and Leib-Mosch,C. (2000). HERV-K-T47D-Related long terminal repeats mediate polyadenylation of cellular transcripts. Genomics *66*, 98-103.

Baust,C., Seifarth,W., Schon,U., Hehlmann,R., and Leib-Mosch,C. (2001). Functional activity of HERV-K-T47D-related long terminal repeats Virology *283*, 262-272.

Beard,K.C., Krishtalka,L., and Stucky,R.K. (1991). First skulls of the early Eocene primate Shoshonius cooperi and the anthropoid-tarsier dichotomy. Nature *349*, 64-67.

Begun,D.R. (2003). Planet of the apes. Sci. Am. *289*, 74-83.

Belshaw,R., Pereira,V., Katzourakis,A., Talbot,G., Paces,J., Burt,A., and Tristem,M. (2004). Long-term reinfection of the human genome by endogenous retroviruses. Proc. Natl. Acad. Sci. U. S. A *101*, 4894-4899.

Benefit,B.R. and McCrossin,M.L. (1997). Earliest known Old World monkey skull. Nature *388*, 368-371.

Benit,L., Dessen,P., and Heidmann,T. (2001). Identification, phylogeny, and evolution of retroviral elements based on their envelope genes J. Virol. *75*, 11709-11719.

Bennett,E.A., Coleman,L.E., Tsui,C., Pittard,W.S., and Devine,S.E. (2004). Natural genetic variation caused by transposable elements in humans. Genetics *168*, 933-951.

Berkhout,B., Jebbink,M., and Zsiros,J. (1999). Identification of an active reverse transcriptase enzyme encoded by a human endogenous HERV-K retrovirus. J. Virol. *73*, 2365-2375.

Bermudez de Castro,J.M., Arsuaga,J.L., Carbonell,E., Rosas,A., Martinez,I., and Mosquera,M. (1997). A hominid from the lower Pleistocene of Atapuerca, Spain: possible ancestor to Neandertals and modern humans. Science *276*, 1392-1395.

Bi,S., Gavrilova,O., Gong,D.W., Mason,M.M., and Reitman,M. (1997). Identification of a placental enhancer for the human leptin gene. J. Biol. Chem. *272*, 30583-30588.

Bieche,I., Laurent,A., Laurendeau,I., Duret,L., Giovangrandi,Y., Frendo,J.L., Olivi,M., Fausser,J.L., Evain-Brion,D., and Vidaud,M. (2003). Placenta-specific INSL4 expression is mediated by a human endogenous retrovirus element. Biol. Reprod. *68*, 1422-1429.

Bieda,K., Hoffmann,A., and Boller,K. (2001). Phenotypic heterogeneity of human endogenous retrovirus particles produced by teratocarcinoma cell lines. J. Gen. Virol. *82*, 591-596.

Blanco,P., Shlumukova,M., Sargent,C.A., Jobling,M.A., Affara,N., and Hurles,M.E. (2000). Divergent outcomes of intrachromosomal recombination on the human Y chromosome: male infertility and recurrent polymorphism J. Med. Genet. *37*, 752-758.

Bloch,J.I. and Silcox,M.T. (2001). New basicrania of Paleocene-Eocene Ignacius: re-evaluation of the Plesiadapiform-Dermopteran link. Am. J. Phys. Anthropol. *116*, 184-198.

Blumenschine,R.J., Peters,C.R., Masao,F.T., Clarke,R.J., Deino,A.L., Hay,R.L., Swisher,C.C., Stanistreet,I.G., Ashley,G.M., McHenry,L.J., Sikes,N.E., Van Der Merwe,N.J., Tactikos,J.C., Cushing,A.E., Deocampo,D.M., Njau,J.K., and Ebert,J.I. (2003). Late Pliocene Homo and hominid land use from Western Olduvai Gorge, Tanzania. Science *299*, 1217-1221.

Boese,A., Galli,U., Geyer,M., Sauter,M., and Mueller-Lantzsch,N. (2001). The Rev/Rex homolog HERV-K cORF multimerizes via a C-terminal domain. FEBS Lett. *493*, 117-121.

Boese,A., Sauter,M., Galli,U., Best,B., Herbst,H., Mayer,J., Kremmer,E., Roemer,K., and Mueller-Lantzsch,N. (2000a). Human endogenous retrovirus protein cORF supports cell transformation and associates with the promyelocytic leukemia zinc finger protein. Oncogene *19*, 4328-4336.

Boese,A., Sauter,M., and Mueller-Lantzsch,N. (2000b). A rev-like NES mediates cytoplasmic localization of HERV-K cORF. FEBS Lett. *468*, 65-67.

Boissinot,S., Chevret,P., and Furano,A.V. (2000). L1 (LINE-1) retrotransposon evolution and amplification in recent human history. Mol. Biol. Evol. *17*, 915-928.

Boissinot,S. and Furano,A.V. (2001). Adaptive evolution in LINE-1 retrotransposons. Mol. Biol. Evol. *18*, 2186-2194.

Boller,K., Janssen,O., Schuldes,H., Tonjes,R.R., and Kurth,R. (1997). Characterization of the antibody response specific for the human endogenous retrovirus HTDV/HERV-K. J. Virol. *71*, 4581-4588.

Boller,K., Konig,H., Sauter,M., Mueller-Lantzsch,N., Lower,R., Lower,J., and Kurth,R. (1993). Evidence that HERV-K is the endogenous retrovirus sequence that codes for the human teratocarcinoma-derived retrovirus HTDV. Virology *196*, 349-353.

Bosch,E. and Jobling,M.A. (2003). Duplications of the AZFa region of the human Y chromosome are mediated by homologous recombination between HERVs and are compatible with male fertility. Hum. Mol. Genet. *12*, 341-347.

Bowler,J.M. and Magee,J.W. (2000). Redating Australia's oldest human remains: a sceptic's view. J. Hum. Evol. *38*, 719-726.

Boyd,M.T., Foley,B., and Brodsky,I. (1997). Evidence for copurification of HERV-K-related transcripts and a reverse transcriptase activity in human platelets from patients with essential thrombocythemia. Blood *90*, 4022-4030.

Britten,R.J. (1994). Evidence that most human Alu sequences were inserted in a process that ceased about 30 million years ago. Proc. Natl. Acad. Sci. U. S. A *91*, 6148-6150.

Britten,R.J. (2002). Divergence between samples of chimpanzee and human DNA sequences is 5%, counting indels. Proc. Natl. Acad. Sci. U. S. A *99*, 13633-13635.

Brosius,J. (1999). Genomes were forged by massive bombardments with retroelements and retrosequences. Genetica *107*, 209-238.

Brouha,B., Schustak,J., Badge,R.M., Lutz-Prigge,S., Farley,A.H., Moran,J.V., and Kazazian,H.H., Jr. (2003). Hot L1s account for the bulk of retrotransposition in the human population. Proc. Natl. Acad. Sci. U. S. A *100*, 5280-5285.

Brown,F., Harris,J., Leakey,R., and Walker,A. (1985). Early Homo erectus skeleton from west Lake Turkana, Kenya. Nature *316*, 788-792.

Brunet,M., Guy,F., Pilbeam,D., Mackaye,H.T., Likius,A., Ahounta,D., Beauvilain,A., Blondel,C., Bocherens,H., Boisserie,J.R., de Bonis,L., Coppens,Y., Dejax,J., Denys,C., Duringer,P., Eisenmann,V., Fanone,G., Fronty,P., Geraads,D.,

Lehmann,T., Lihoreau,F., Louchart,A., Mahamat,A., Merceron,G., Mouchelin,G., Otero,O., Pelaez,C.P., Ponce,D.L., Rage,J.C., Sapanet,M., Schuster,M., Sudre,J., Tassy,P., Valentin,X., Vignaud,P., Viriot,L., Zazzo,A., and Zollikofer,C. (2002). A new hominid from the Upper Miocene of Chad, Central Africa. Nature *418*, 145-151.

Buzdin,A., Khodosevich,K., Mamedov,I., Vinogradova,T., Lebedev,Y., Hunsmann,G., and Sverdlov,E. (2002). A technique for genome-wide identification of differences in the interspersed repeats integrations between closely related genomes and its application to detection of human-specific integrations of HERV-K LTRs. Genomics *79*, 413-422.

Buzdin,A., Ustyugova,S., Khodosevich,K., Mamedov,I., Lebedev,Y., Hunsmann,G., and Sverdlov,E. (2003). Human-specific subfamilies of HERV-K (HML-2) long terminal repeats: three master genes were active simultaneously during branching of hominoid lineages (small star, filled)Genomics *81*, 149-156.

Callahan,R., Chiu,I.M., Wong,J.F., Tronick,S.R., Roe,B.A., Aaronson,S.A., and Schlom,J. (1985). A new class of endogenous human retroviral genomes. Science *228*, 1208-1211.

Callahan,R., Drohan,W., Tronick,S., and Schlom,J. (1982). Detection and cloning of human DNA sequences related to the mouse mammary tumor virus genome. Proc. Natl. Acad. Sci. U. S. A *79*, 5503-5507.

Cann,R.L. (2002). Human evolution: tangled genetic routes Nature *416*, 32-33.

Cann,R.L., Stoneking,M., and Wilson,A.C. (1987). Mitochondrial DNA and human evolution. Nature *325*, 31-36.

Caramelli,D., Lalueza-Fox,C., Vernesi,C., Lari,M., Casoli,A., Mallegni,F., Chiarelli,B., Dupanloup,I., Bertranpetit,J., Barbujani,G., and Bertorelle,G. (2003). Evidence for a genetic discontinuity between Neandertals and 24,000-year-old anatomically modern Europeans. Proc. Natl. Acad. Sci. U. S. A *100*, 6593-6597.

Carbonell,E., Esteban,M., Najera,A.M., Mosquera,M., Rodriguez,X.P., Olle,A., Sala,R., Verges,J.M., Bermudez de Castro,J.M., and Ortega,A.I. (1999). The Pleistocene site of Gran Dolina, Sierra de Atapuerca, Spain: a history of the archaeological investigations. J. Hum. Evol. *37*, 313-324.

Carretero,J.M., Lorenzo,C., and Arsuaga,J.L. (1999). Axial and appendicular skeleton of Homo antecessor. J. Hum. Evol. *37*, 459-499.

Cavalli-Sforza,L.L., Menozzi,P., and Piazza,A. (1994). The History and Geography of Human Genes. Princeton University Press).

Chakravarti,A. (1999). Population genetics--making sense out of sequence. Nat. Genet. *21*, 56-60.

Chen,F.C. and Li,W.H. (2001). Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. Am. J. Hum. Genet. *68*, 444-456.

Ciochon,R., Long,V.T., Larick,R., Gonzalez,L., Grun,R., de Vos,J., Yonge,C., Taylor,L., Yoshida,H., and Reagan,M. (1996). Dated co-occurrence of Homo erectus and Gigantopithecus from Tham Khuyen Cave, Vietnam. Proc. Natl. Acad. Sci. U. S. A *93*, 3016-3020.

Conrad,B., Weissmahr,R.N., Boni,J., Arcari,R., Schupbach,J., and Mach,B. (1997). A human endogenous retroviral superantigen as candidate autoimmune gene in type I diabetes. Cell *90*, 303-313.

Cooper,A., Rambaut,A., Macaulay,V., Willerslev,E., Hansen,A.J., and Stringer,C. (2001). Human origins and ancient human DNA. Science *292*, 1655-1656.

Costas,J. (2001). Evolutionary dynamics of the human endogenous retrovirus family HERV-K inferred from full-length proviral genomes. J. Mol. Evol. *53*, 237-243.

Costas,J. (2002). Characterization of the intragenomic spread of the human endogenous retrovirus family HERV-W. Mol. Biol. Evol. *19*, 526-533.

Costas,J. and Naveira,H. (2000). Evolutionary history of the human endogenous retrovirus family ERV9. Mol. Biol. Evol. *17*, 320-330.

Curnoe,D. and Thorne,A. (2003). Number of ancestral human species: a molecular perspective. Homo. *53*, 201-224.

Dangel,A.W., Baker,B.J., Mendoza,A.R., and Yu,C.Y. (1995). Complement component C4 gene intron 9 as a phylogenetic marker for primates: long terminal repeats of the endogenous retrovirus ERV-K(C4) are a molecular clock of evolution Immunogenetics *42*, 41-52.

De Bonis,L., Bouvrain,G., Geraads,D., and Koufos,G. (1990). New hominid skull material from the late Miocene of Macedonia in northern Greece. Nature *345*, 712-714.

De Parseval,N., Lazar,V., Casella,J.F., Benit,L., and Heidmann,T. (2003). Survey of human genes of retroviral origin: identification and transcriptome of the genes with coding capacity for complete envelope proteins. J. Virol. *77*, 10414-10422.

Deacon,H.J. (1992). Southern Africa and modern human origins. Philos. Trans. R. Soc. Lond B Biol. Sci. *337*, 177-183.

Deen,K.C. and Sweet,R.W. (1986). Murine mammary tumor virus pol-related sequences in human DNA: characterization and sequence comparison with the complete murine mammary tumor virus pol gene. J. Virol. *57* , 422-432.

Deininger,P.L. and Batzer,M.A. (1999). Alu repeats and human disease. Mol. Genet. Metab *67*, 183-193.

Deininger,P.L. and Batzer,M.A. (2002). Mammalian retroelements. Genome Res. *12*, 1455-1465.

Deininger,P.L., Moran,J.V., Batzer,M.A., and Kazazian,H.H., Jr. (2003). Mobile elements and mammalian genome evolution. Curr. Opin. Genet. Dev. *13*, 651-658.

Domansky,A.N., Kopantzev,E.P., Snezhkov,E.V., Lebedev,Y.B., Leib-Mosch,C., and Sverdlov,E.D. (2000). Solitary HERV-K LTRs possess bi-directional promoter activity and contain a negative regulatory element in the U5 region. FEBS Lett. *472*, 191-195.

Duarte,C., Mauricio,J., Pettitt,P.B., Souto,P., Trinkaus,E., van der,P.H., and Zilhao,J. (1999). The early Upper Paleolithic human skeleton from the Abrigo do Lagar Velho (Portugal) and modern human emergence in Iberia. Proc. Natl. Acad. Sci. U. S. A *96*, 7604-7609.

Eichler,E.E. (2001). Recent duplication, domain accretion and the dynamic mutation of the human genome. Trends Genet. *17*, 661-669.

Eichler,E.E., Johnson,M.E., Alkan,C., Tuzun,E., Sahinalp,C., Misceo,D., Archidiacono,N., and Rocchi,M. (2001). Divergent origins and concerted expansion of two segmental duplications on chromosome 16. J. Hered. *92*, 462-468.

Eyre-Walker,A. and Awadalla,P. (2001). Does human mtDNA recombine? J. Mol. Evol. *53*, 430-435.

Faerman,M., Filon,D., Kahila,G., Greenblatt,C.L., Smith,P., and Oppenheim,A. (1995). Sex identification of archaeological human remains based on amplification of the X and Y amelogenin alleles. Gene *167*, 327-332.

Felsenstein,J. (1985). Confidence limits on phylogenies: An approach to using the bootstrap. Evolution *39*, 783-791.

Felsenstein, J. PHYLIP. *Phylogenetic Inference Package.* [Version 3.5]. 1993. Department of Genetics, University of Washington. (Computer Program)

Fleagle,J.G. and Simons,E.L. (1982). The humerus of Aegyptopithecus zeuxis: a primitive anthropoid. Am. J. Phys. Anthropol. *59*, 175-193.

Fortna,A., Kim,Y., MacLaren,E., Marshall,K., Hahn,G., Meltesen,L., Brenton,M., Hink,R., Burgers,S., Hernandez-Boussard,T., Karimpour-Fard,A., Glueck,D.,

McGavran,L., Berry,R., Pollack,J., and Sikela,J.M. (2004). Lineage-specific gene duplication and loss in human and great ape evolution. PLoS. Biol. *2*, E207.

Franklin,G.C., Chretien,S., Hanson,I.M., Rochefort,H., May,F.E., and Westley,B.R. (1988). Expression of human sequences related to those of mouse mammary tumor virus. J. Virol. *62*, 1203-1210.

Frazer,K.A., Chen,X., Hinds,D.A., Pant,P.V., Patil,N., and Cox,D.R. (2003). Genomic DNA insertions and deletions occur frequently between humans and nonhuman primates. Genome Res. *13*, 341-346.

Fujiyama,A., Watanabe,H., Toyoda,A., Taylor,T.D., Itoh,T., Tsai,S.F., Park,H.S., Yaspo,M.L., Lehrach,H., Chen,Z., Fu,G., Saitou,N., Osoegawa,K., de Jong,P.J., Suto,Y., Hattori,M., and Sakaki,Y. (2002). Construction and analysis of a human-chimpanzee comparative clone map. Science *295* , 131-134.

Gabunia,L., Vekua,A., Lordkipanidze,D., Swisher,C.C., III, Ferring,R., Justus,A., Nioradze,M., Tvalchrelidze,M., Anton,S.C., Bosinski,G., Joris,O., Lumley,M.A., Majsuradze,G., and Mouskhelishvili,A. (2000). Earliest Pleistocene hominid cranial remains from Dmanisi, Republic of Georgia: taxonomy, geological setting, and age. Science *288*, 1019-1025.

Gagneux,P. and Varki,A. (2001). Genetic differences between humans and great apes. Mol. Phylogenet. Evol. *18*, 2-13.

Gebo,D.L., Dagosto,M., Beard,K.C., Qi,T., and Wang,J. (2000). The oldest known anthropoid postcranial fossils and the early evolution of higher primates. Nature *404*, 276-278.

Gilbert,M.T., Willerslev,E., Hansen,A.J., Barnes,I., Rudbeck,L., Lynnerup,N., and Cooper,A. (2003). Distribution patterns of postmortem damage in human mitochondrial DNA. Am. J. Hum. Genet. *72*, 32-47.

Glazko,G.V. and Nei,M. (2003). Estimation of divergence times for major lineages of primate species. Mol. Biol. Evol. *20*, 424-434.

Goff,S.P. (2001). The Retroviruses and Their Replication. In Fields Virology, D.Knipe and P.Howley, eds. (Lippincott Williams and Wilkins), pp. 1871-1939.

Goodchild,N.L., Freeman,J.D., and Mager,D.L. (1995). Spliced HERV-H endogenous retroviral sequences in human genomic DNA: evidence for amplification via retrotransposition. Virology *206*, 164-173.

Goodman,M., Bailey,W.J., Hayasaka,K., Stanhope,M.J., Slightom,J., and Czelusniak,J. (1994). Molecular evidence on primate phylogeny from DNA sequences. Am. J. Phys. Anthropol. *94*, 3-24.

Gotzinger,N., Sauter,M., Roemer,K., and Mueller-Lantzsch,N. (1996). Regulation of human endogenous retrovirus-K Gag expression in teratocarcinoma cell lines and human tumours. J. Gen. Virol. *77*, 2983-2990.

Greenwood,A.D., Lee,F., Capelli,C., DeSalle,R., Tikhonov,A., Marx,P.A., and MacPhee,R.D. (2001). Evolution of endogenous retrovirus-like elements of the woolly mammoth (Mammuthus primigenius) and its relatives. Mol. Biol. Evol. *18*, 840-847.

Griffiths,D.J. (2001). Endogenous retroviruses in the human genome sequence. Genome Biol. *2*, REVIEWS1017.

Grun,R. and Thorne,A. (1997). Dating the Ngandong humans. Science *276*, 1575-1576.

Haig,D. (1999). A brief history of human autosomes. Philos. Trans. R. Soc. Lond B Biol. Sci. *354*, 1447-1470.

Haile-Selassie,Y., Asfaw,B., and White,T.D. (2004). Hominid cranial remains from upper Pleistocene deposits at Aduma, Middle Awash, Ethiopia. Am. J. Phys. Anthropol. *123*, 1-10.

Hammer,M.F., Karafet,T., Rasanayagam,A., Wood,E.T., Altheide,T.K., Jenkins,T., Griffiths,R.C., Templeton,A.R., and Zegura,S.L. (1998). Out of Africa and back

again: nested cladistic analysis of human Y chromosome variation Mol. Biol. Evol. *15*, 427-441.

Hardies,S.C., Martin,S.L., Voliva,C.F., Hutchison,C.A., III, and Edgell,M.H. (1986). An analysis of replacement and synonymous changes in the rodent L1 repeat family. Mol. Biol. Evol. *3*, 109-125.

Harding,R.M., Fullerton,S.M., Griffiths,R.C., Bond,J., Cox,M.J., Schneider,J.A., Moulin,D.S., and Clegg,J.B. (1997). Archaic African and Asian lineages in the genetic ancestry of modern humans. Am. J. Hum. Genet. *60*, 772-789.

Harpending,H. and Rogers,A. (2000). Genetic perspectives on human origins and differentiation. Annu. Rev. Genomics Hum. Genet. *1*, 361-385.

Harris,E.E. and Hey,J. (1999). X chromosome evidence for ancient human histories Proc. Natl. Acad. Sci. U. S. A *96*, 3320-3324.

Hasegawa,M., Kishino,H., and Yano,T. (1987). Man's place in Hominoidea as inferred from molecular clocks of DNA. J. Mol. Evol. *26*, 132-147.

Hedges,S.B., Parker,P.H., Sibley,C.G., and Kumar,S. (1996). Continental breakup and the ordinal diversification of birds and mammals. Nature *381*, 226-229.

Herbst,H., Sauter,M., Kuhler-Obbarius,C., Loning,T., and Mueller-Lantzsch,N. (1998). Human endogenous retrovirus (HERV)-K transcripts in germ cell and trophoblastic tumours. APMIS *106*, 216-220.

Herbst,H., Sauter,M., and Mueller-Lantzsch,N. (1996). Expression of human endogenous retrovirus K elements in germ cell and trophoblastic tumors. Am. J. Pathol. *149*, 1727-1735.

Horton,R., Niblett,D., Milne,S., Palmer,S., Tubby,B., Trowsdale,J., and Beck,S. (1998). Large-scale sequence comparisons reveal unusually high levels of variation in the HLA-DQB1 locus in the class II region of the human MHC J. Mol. Biol. *282*, 71-97.

Huang,W., Fu,Y.X., Chang,B.H., Gu,X., Jorde,L.B., and Li,W.H. (1998). Sequence variation in ZFX introns in human populations. Mol. Biol. Evol. *15*, 138-142.

Hughes,J.F. and Coffin,J.M. (2001). Evidence for genomic rearrangements mediated by human endogenous retroviruses during primate evolution. Nat. Genet. *29*, 487-489.

Hughes,J.F. and Coffin,J.M. (2004). Human endogenous retrovirus K solo-LTR formation and insertional polymorphisms: implications for human and viral evolution. Proc. Natl. Acad. Sci. U. S. A *101*, 1668-1672.

Huh,J.W., Hong,K.W., Yi,J.M., Kim,T.H., Takenaka,O., Lee,W.H., and Kim,H.S. (2003). Molecular phylogeny and evolution of the human endogenous retrovirus HERV-W LTR family in hominoid primates. Mol. Cells *15*, 122-126.

Hurles,M.E. and Jobling,M.A. (2003). A singular chromosome. Nat. Genet. *34*, 246-247.

Hurles,M.E., Willey,D., Matthews,L., and Hussain,S.S. (2004). Origins of chromosomal rearrangement hotspots in the human genome: evidence from the AZFa deletion hotspots. Genome Biol. *5*, R55.

Jaeger,J., Thein,T., Benammi,M., Chaimanee,Y., Soe,A.N., Lwin,T., Tun,T., Wai,S., and Ducrocq,S. (1999). A new primate from the Middle Eocene of Myanmar and the Asian early origin of anthropoids. Science *286*, 528-530.

Jeffreys,A.J. and May,C.A. (2004). Intense and highly localized gene conversion activity in human meiotic crossover hot spots. Nat. Genet. *36*, 151-156.

Jeffreys,A.J., Neil,D.L., and Neumann,R. (1998). Repeat instability at human minisatellites arising from meiotic recombination. EMBO J. *17*, 4147-4157.

Jeffreys,A.J. and Neumann,R. (2002). Reciprocal crossover asymmetry and meiotic drive in a human recombination hot spot. Nat. Genet. *31*, 267-271.

Johnson,M.E., Viggiano,L., Bailey,J.A., Abdul-Rauf,M., Goodwin,G., Rocchi,M., and Eichler,E.E. (2001). Positive selection of a gene family during the emergence of humans and African apes. Nature *413*, 514-519.

Johnson,W.E. and Coffin,J.M. (1999). Constructing primate phylogenies from ancient retrovirus sequences. Proc. Natl. Acad. Sci. U. S. A *96*, 10254-10260.

Jorde,L.B., Bamshad,M.J., Watkins,W.S., Zenger,R., Fraley,A.E., Krakowiak,P.A., Carpenter,K.D., Soodyall,H., Jenkins,T., and Rogers,A.R. (1995). Origins and affinities of modern humans: a comparison of mitochondrial and nuclear genetic data. Am. J. Hum. Genet. *57*, 523-538.

Jorde,L.B., Watkins,W.S., Bamshad,M.J., Dixon,M.E., Ricker,C.E., Seielstad,M.T., and Batzer,M.A. (2000). The distribution of human genetic diversity: a comparison of mitochondrial, autosomal, and Y-chromosome data. Am. J. Hum. Genet. *66*, 979-988.

Jukes,T. and Cantor,C.R. (1969). Evolution of protein molecules. In Mammalian Protein Metabolism, H.N.Munro, ed. New York, Academic Press), pp. 21-132.

Jurka,J. (2000). Repbase update: a database and an electronic journal of repetitive elements. Trends Genet. *16*, 418-420.

Kaessmann,H., Heissig,F., von Haeseler,A., and Paabo,S. (1999). DNA sequence variation in a non-coding region of low recombination on the human X chromosome Nat. Genet. *22*, 78-81.

Kaessmann,H., Wiebe,V., Weiss,G., and Paabo,S. (2001). Great ape DNA sequences reveal a reduced diversity and an expansion in humans. Nat. Genet. *27*, 155-156.

Kamp,C., Hirschmann,P., Voss,H., Huellen,K., and Vogt,P.H. (2000). Two long homologous retroviral sequence blocks in proximal Yq11 cause AZFa microdeletions as a result of intrachromosomal recombination events. Hum. Mol. Genet. *9*, 2563-2572.

Kapitonov,V.V. and Jurka,J. (1999). The long terminal repeat of an endogenous retrovirus induces alternative splicing and encodes an additional carboxy-terminal sequence in the human leptin receptor. J. Mol. Evol. *48*, 248-251.

Keller,G., Adatte,T., Stinnesbeck,W., Rebolledo-Vieyra,M., Fucugauchi,J.U., Kramar,U., and Stuben,D. (2004). Chicxulub impact predates the K-T boundary mass extinction. Proc. Natl. Acad. Sci. U. S. A *101*, 3753-3758.

Kim,A., Jun,H.S., Wong,L., Stephure,D., Pacaud,D., Trussell,R.A., and Yoon,J.W. (1999a). Human endogenous retrovirus with a high genomic sequence homology with IDDMK(1,2)22 is not specific for Type I (insulin-dependent) diabetic patients but ubiquitous. Diabetologia *42*, 413-418.

Kim,H.S., Hyun,B.H., and Crow,T.J. (2000). Phylogenetic analysis of retroposon family as exemplified on human chromosome 13: further evidence for recent proliferation. Mol. Cells *10*, 356-360.

Kim,H.S., Hyun,B.H., and Takenaka,O. (2002). Isolation and phylogeny of endogenous retrovirus HERV-F family in Old World monkeys. Brief report. Arch. Virol. *147*, 393-400.

Kim,H.S., Takenaka,O., and Crow,T.J. (1999b). Cloning and nucleotide sequence of retroposons specific to hominoid primates derived from an endogenous retrovirus (HERV-K) AIDS Res. Hum. Retroviruses *15*, 595-601.

Kim,H.S., Takenaka,O., and Crow,T.J. (1999c). Isolation and phylogeny of endogenous retrovirus sequences belonging to the HERV-W family in primates J. Gen. Virol. *80 (Pt 10)*, 2613-2619.

Kimura,M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. *16*, 111-120.

King,M.C. and Wilson,A.C. (1975). Evolution at two levels in humans and chimpanzees. Science *188*, 107-116.

Kitamura,Y., Ayukawa,T., Ishikawa,T., Kanda,T., and Yoshiike,K. (1996). Human endogenous retrovirus K10 encodes a functional integrase. J. Virol. *70*, 3302-3306.

Kolomietz,E., Meyn,M.S., Pandita,A., and Squire,J.A. (2002). The role of Alu repeat clusters as mediators of recurrent chromosomal aberrations in tumors. Genes Chromosomes. Cancer *35*, 97-112.

Kramer,A. (1993). Human taxonomic diversity in the pleistocene: does Homo erectus represent multiple hominid species? Am. J. Phys. Anthropol. *91*, 161-171.

Krings,M., Capelli,C., Tschentscher,F., Geisert,H., Meyer,S., von Haeseler,A., Grossschmidt,K., Possnert,G., Paunovic,M., and Paabo,S. (2000). A view of Neandertal genetic diversity. Nat. Genet. *26* , 144-146.

Krings,M., Geisert,H., Schmitz,R.W., Krainitzki,H., and Paabo,S. (1999). DNA sequence of the mitochondrial hypervariable region II from the neandertal type specimen. Proc. Natl. Acad. Sci. U. S. A *96*, 5581-5585.

Krings,M., Stone,A., Schmitz,R.W., Krainitzki,H., Stoneking,M., and Paabo,S. (1997). Neandertal DNA sequences and the origin of modern humans. Cell *90*, 19-30.

Kulski,J.K., Gaudieri,S., Inoko,H., and Dawkins,R.L. (1999a). Comparison between two human endogenous retrovirus (HERV)-rich regions within the major histocompatibility complex. J. Mol. Evol. *48*, 675-683.

Kulski,J.K., Gaudieri,S., Martin,A., and Dawkins,R.L. (1999b). Coevolution of PERB11 (MIC) and HLA class I genes with HERV-16 and retroelements by extended genomic duplication. J. Mol. Evol. *49*, 84-97.

Kumar,S. and Hedges,S.B. (1998). A molecular timescale for vertebrate evolution. Nature *392*, 917-920.

Kumar, S., Tamura, K., Jakobsen I, B, and Nei, M. MEGA2: Molecular Evolutionary Genetics Analysis software. [Version 2.1]. 2001. Arizona State University, Tempe, Arizona, USA. (Computer Program)

Kumar,S. and Subramanian,S. (2002). Mutation rates in mammalian genomes. Proc. Natl. Acad. Sci. U. S. A *99*, 803-808.

Kurdyukov,S.G., Lebedev,Y.B., Artamonova,I.I., Gorodentseva,T.N., Batrak,A.V., Mamedov,I.Z., Azhikina,T.L., Legchilina,S.P., Efimenko,I.G., Gardiner,K., and Sverdlov,E.D. (2001). Full-sized HERV-K (HML-2) human endogenous retroviral LTR sequences on human chromosome 21: map locations and evolutionary history. Gene *273*, 51-61.

Lander,E.S., Linton,L.M., Birren,B., Nusbaum,C., Zody,M.C., Baldwin,J., Devon,K., Dewar,K., Doyle,M., FitzHugh,W., Funke,R., Gage,D., Harris,K., Heaford,A., Howland,J., Kann,L., Lehoczky,J., LeVine,R., McEwan,P., McKernan,K., Meldrim,J., Mesirov,J.P., Miranda,C., Morris,W., Naylor,J., Raymond,C., Rosetti,M., Santos,R., Sheridan,A., Sougnez,C., Stange-Thomann,N., Stojanovic,N., Subramanian,A., Wyman,D., Rogers,J., Sulston,J., Ainscough,R., Beck,S., Bentley,D., Burton,J., Clee,C., Carter,N., Coulson,A., Deadman,R., Deloukas,P., Dunham,A., Dunham,I., Durbin,R., French,L., Grafham,D., Gregory,S., Hubbard,T., Humphray,S., Hunt,A., Jones,M., Lloyd,C., McMurray,A., Matthews,L., Mercer,S., Milne,S., Mullikin,J.C., Mungall,A., Plumb,R., Ross,M., Shownkeen,R., Sims,S., Waterston,R.H., Wilson,R.K., Hillier,L.W., McPherson,J.D., Marra,M.A., Mardis,E.R., Fulton,L.A., Chinwalla,A.T., Pepin,K.H., Gish,W.R., Chissoe,S.L., Wendl,M.C., Delehaunty,K.D., Miner,T.L., Delehaunty,A., Kramer,J.B., Cook,L.L., Fulton,R.S., Johnson,D.L., Minx,P.J., Clifton,S.W., Hawkins,T., Branscomb,E., Predki,P., Richardson,P., Wenning,S., Slezak,T., Doggett,N., Cheng,J.F., Olsen,A., Lucas,S., Elkin,C., Uberbacher,E., Frazier,M., Gibbs,R.A., Muzny,D.M., Scherer,S.E., Bouck,J.B., Sodergren,E.J., Worley,K.C., Rives,C.M., Gorrell,J.H., Metzker,M.L., Naylor,S.L., Kucherlapati,R.S., Nelson,D.L., Weinstock,G.M., Sakaki,Y., Fujiyama,A., Hattori,M., Yada,T., Toyoda,A., Itoh,T., Kawagoe,C., Watanabe,H., Totoki,Y., Taylor,T., Weissenbach,J., Heilig,R., Saurin,W., Artiguenave,F., Brottier,P., Bruls,T., Pelletier,E., Robert,C., Wincker,P., Smith,D.R., Doucette-Stamm,L., Rubenfield,M., Weinstock,K., Lee,H.M., Dubois,J., Rosenthal,A., Platzer,M., Nyakatura,G., Taudien,S., Rump,A., Yang,H., Yu,J., Wang,J., Huang,G., Gu,J., Hood,L., Rowen,L., Madan,A., Qin,S., Davis,R.W.,

Federspiel,N.A., Abola,A.P., Proctor,M.J., Myers,R.M., Schmutz,J., Dickson,M., Grimwood,J., Cox,D.R., Olson,M.V., Kaul,R., Raymond,C., Shimizu,N., Kawasaki,K., Minoshima,S., Evans,G.A., Athanasiou,M., Schultz,R., Roe,B.A., Chen,F., Pan,H., Ramser,J., Lehrach,H., Reinhardt,R., McCombie,W.R., de la,B.M., Dedhia,N., Blocker,H., Hornischer,K., Nordsiek,G., Agarwala,R., Aravind,L., Bailey,J.A., Bateman,A., Batzoglou,S., Birney,E., Bork,P., Brown,D.G., Burge,C.B., Cerutti,L., Chen,H.C., Church,D., Clamp,M., Copley,R.R., Doerks,T., Eddy,S.R., Eichler,E.E., Furey,T.S., Galagan,J., Gilbert,J.G., Harmon,C., Hayashizaki,Y., Haussler,D., Hermjakob,H., Hokamp,K., Jang,W., Johnson,L.S., Jones,T.A., Kasif,S., Kaspryzk,A., Kennedy,S., Kent,W.J., Kitts,P., Koonin,E.V., Korf,I., Kulp,D., Lancet,D., Lowe,T.M., McLysaght,A., Mikkelsen,T., Moran,J.V., Mulder,N., Pollara,V.J., Ponting,C.P., Schuler,G., Schultz,J., Slater,G., Smit,A.F., Stupka,E., Szustakowski,J., Thierry-Mieg,D., Thierry-Mieg,J., Wagner,L., Wallis,J., Wheeler,R., Williams,A., Wolf,Y.I., Wolfe,K.H., Yang,S.P., Yeh,R.F., Collins,F., Guyer,M.S., Peterson,J., Felsenfeld,A., Wetterstrand,K.A., Patrinos,A., Morgan,M.J., Szustakowki,J., de Jong,P., Catanese,J.J., Osoegawa,K., Shizuya,H., and Choi,S. (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860-921.

Lapuk,A.V., Khil,P.P., Lavrentieva,I.V., Lebedev,Y.B., and Sverdlov,E.D. (1999). A human endogenous retrovirus-like (HERV) LTR formed more than 10 million years ago due to an insertion of HERV-H LTR into the 5' LTR of HERV-K is situated on human chromosomes 10, 19 and YJ. Gen. Virol. *80 (Pt 4)*, 835-839.

Larsson,E., Kato,N., and Cohen,M. (1989). Human endogenous proviruses. Curr. Top. Microbiol. Immunol. *148:115-32.*, 115-132.

Lavie,L., Medstrand,P., Schempp,W., Meese,E., and Mayer,J. (2004). Human endogenous retrovirus family HERV-K(HML-5): status, evolution, and reconstruction of an ancient betaretrovirus in the human genome. J. Virol. *78*, 8788-8798.

Lavrentieva,I., Khil,P., Vinogradova,T., Akhmedov,A., Lapuk,A., Shakhova,O., Lebedev,Y., Monastyrskaya,G., and Sverdlov,E.D. (1998). Subfamilies and nearest-neighbour dendrogram for the LTRs of human endogenous retroviruses HERV-K

mapped on human chromosome 19: physical neighbourhood does not correlate with identity level. Hum. Genet. *102*, 107-116.

Leakey,M.G., Feibel,C.S., McDougall,I., and Walker,A. (1995). New four-million-year-old hominid species from Kanapoi and Allia Bay, Kenya. Nature *376*, 565-571.

Leakey,M.G., Spoor,F., Brown,F.H., Gathogo,P.N., Kiarie,C., Leakey,L.N., and McDougall,I. (2001). New hominin genus from eastern Africa shows diverse middle Pliocene lineages. Nature *410*, 433-440.

Lebedev,Y.B., Belonovitch,O.S., Zybrova,N.V., Khil,P.P., Kurdyukov,S.G., Vinogradova,T.V., Hunsmann,G., and Sverdlov,E.D. (2000). Differences in HERV-K LTR insertions in orthologous loci of humans and great apes. Gene *247*, 265-277.

Leib-Mosch,C., Haltmeier,M., Werner,T., Geigl,E.M., Brack-Werner,R., Francke,U., Erfle,V., and Hehlmann,R. (1993). Genomic distribution and transcription of solitary HERV-K LTRs. Genomics *18*, 261-269.

Leib-Mosch,C. and Seifarth,W. (1995). Evolution and biological significance of human retroelements. Virus Genes *11*, 133-145.

Levene,H. (1949). On a matching problem in genetics. Annuals of Mathematical Statistics *20*, 91-94.

Li,M.D., Bronson,D.L., Lemke,T.D., and Faras,A.J. (1995). Restricted expression of new HERV-K members in human teratocarcinoma cells. Virology *20;208*, 733-741.

Liao,D., Pavelitz,T., and Weiner,A.M. (1998). Characterization of a novel class of interspersed LTR elements in primate genomes: structure, genomic distribution, and evolution J. Mol. Evol. *46*, 649-660.

Locke,D.P., Segraves,R., Carbone,L., Archidiacono,N., Albertson,D.G., Pinkel,D., and Eichler,E.E. (2003). Large-scale variation among human and great ape genomes determined by array comparative genomic hybridization. Genome Res. *13*, 347-357.

Lower,R., Boller,K., Hasenmaier,B., Korbmacher,C., Muller-Lantzsch,N., Lower,J., and Kurth,R. (1993a). Identification of human endogenous retroviruses with complex mRNA expression and particle formation. Proc. Natl. Acad. Sci. U. S. A *90*, 4480-4484.

Lower,R., Lower,J., and Kurth,R. (1996). The viruses in all of us: characteristics and biological significance of human endogenous retrovirus sequences. Proc. Natl. Acad. Sci. U. S. A *93*, 5177-5184.

Lower,R., Lower,J., Tondera-Koch,C., and Kurth,R. (1993b). A general method for the identification of transcribed retrovirus sequences (R-U5 PCR) reveals the expression of the human endogenous retrovirus loci HERV-H and HERV-K in teratocarcinoma cells. Virology *192*, 501-511.

Lower,R., Tonjes,R.R., Boller,K., Denner,J., Kaiser,B., Phelps,R.C., Lower,J., Kurth,R., Badenhoop,K., Donner,H., Usadel,K.H., Miethke,T., Lapatschek,M., and Wagner,H. (1998). Development of insulin-dependent diabetes mellitus does not depend on specific expression of the human endogenous retrovirus HERV-K. Cell *95*, 11-14.

Lower,R., Tonjes,R.R., Korbmacher,C., Kurth,R., and Lower,J. (1995). Identification of a Rev-related protein by analysis of spliced transcripts of the human endogenous retroviruses HTDV/HERV-K. J. Virol. *69*, 141-149.

Lutz,S.M., Vincent,B.J., Kazazian,H.H., Jr., Batzer,M.A., and Moran,J.V. (2003). Allelic heterogeneity in LINE-1 retrotransposition activity. Am. J. Hum. Genet. *73*, 1431-1437.

Macfarlane,C. and Simmonds,P. (2004). Allelic variation of HERV-K(HML-2) Endogenous Retroviral elements in Human Populations. J. Mol. Evol. *59*, 642-656.

Madar,S.I., Rose,M.D., Kelley,J., MacLatchy,L., and Pilbeam,D. (2002). New Sivapithecus postcranial specimens from the Siwaliks of Pakistan. J. Hum. Evol. *42*, 705-752.

Mager,D.L. and Freeman,J.D. (1995). HERV-H endogenous retroviruses: presence in the New World branch but amplification in the Old World primate lineage Virology *213*, 395-404.

Mager,D.L. and Goodchild,N.L. (1989). Homologous recombination between the LTRs of a human retrovirus-like element causes a 5-kb deletion in two siblings Am. J. Hum. Genet. *45*, 848-854.

Magin-Lachmann,C., Hahn,S., Strobel,H., Held,U., Lower,J., and Lower,R. (2001). Rec (formerly Corf) function requires interaction with a complex, folded RNA structure within its responsive element rather than binding to a discrete specific binding site. J. Virol. *75*, 10359-10371.

Mamedov,I., Batrak,A., Buzdin,A., Arzumanyan,E., Lebedev,Y., and Sverdlov,E.D. (2002). Genome-wide comparison of differences in the integration sites of interspersed repeats between closely related genomes. Nucleic Acids Res. *30*, e71.

Manzi,G., Mallegni,F., and Ascenzi,A. (2001). A cranium for the earliest Europeans: phylogenetic position of the hominid from Ceprano, Italy. Proc. Natl. Acad. Sci. U. S. A *98*, 10011-10016.

Mariani-Costantini,R., Horn,T.M., and Callahan,R. (1989). Ancestry of a human endogenous retrovirus family. J. Virol. *63*, 4982-4985.

Mayer,J., Ehlhardt,S., Seifert,M., Sauter,M., Muller-Lantzsch,N., Mehraein,Y., Zang,K.D., and Meese,E. (2004). Human endogenous retrovirus HERV-K(HML-2) proviruses with Rec protein coding capacity and transcriptional activity. Virology *322*, 190-198.

Mayer,J., Meese,E., and Mueller-Lantzsch,N. (1997a). Chromosomal assignment of human endogenous retrovirus K (HERV-K) env open reading frames. Cytogenet. Cell Genet. *79*, 157-161.

Mayer,J., Meese,E., and Mueller-Lantzsch,N. (1997b). Multiple human endogenous retrovirus (HERV-K) loci with gag open reading frames in the human genome. Cytogenet. Cell Genet. *78*, 1-5.

Mayer,J., Meese,E., and Mueller-Lantzsch,N. (1998). Human endogenous retrovirus K homologous sequences and their coding capacity in Old World primates. J. Virol. *72*, 1870-1875.

Mayer,J. and Meese,E.U. (2002). The human endogenous retrovirus family HERV-K(HML-3). Genomics *80*, 331-343.

Mayer,J., Sauter,M., Racz,A., Scherer,D., Mueller-Lantzsch,N., and Meese,E. (1999). An almost-intact human endogenous retrovirus K on human chromosome 7. Nat. Genet. *21*, 257-258.

Mayer,W.E., O'Huigin,C., and Klein,J. (1993). Resolution of the HLA-DRB6 puzzle: a case of grafting a de novo-generated exon on an existing gene. Proc. Natl. Acad. Sci. U. S. A *90*, 10720-10724.

McCrossin,M.L. and Benefit,B.R. (1993). Recently recovered Kenyapithecus mandible and its implications for great ape and human origins. Proc. Natl. Acad. Sci. U. S. A *90*, 1962-1966.

Medstrand,P. and Blomberg,J. (1993). Characterization of novel reverse transcriptase encoding human endogenous retroviral sequences similar to type A and type B retroviruses: differential transcription in normal human tissues. J. Virol. *67*, 6778-6787.

Medstrand,P., Lindeskog,M., and Blomberg,J. (1992). Expression of human endogenous retroviral sequences in peripheral blood mononuclear cells of healthy individuals. J. Gen. Virol. *73*, 2463-2466.

Medstrand,P. and Mager,D.L. (1998). Human-specific integrations of the HERV-K endogenous retrovirus family. J. Virol. *72*, 9782-9787.

Medstrand,P., Mager,D.L., Yin,H., Dietrich,U., and Blomberg,J. (1997). Structure and genomic organization of a novel human endogenous retrovirus family: HERV-K (HML-6). J. Gen. Virol. *78*, 1731-1744.

Mefford,H.C., Linardopoulou,E., Coil,D., van den,E.G., and Trask,B.J. (2001). Comparative sequencing of a multicopy subtelomeric region containing olfactory receptor genes reveals multiple interactions between non-homologous chromosomes. Hum. Mol. Genet. *10*, 2363-2372.

Minghetti,P.P. and Dugaiczyk,A. (1993). The emergence of new DNA repeats and the divergence of primates. Proc. Natl. Acad. Sci. U. S. A *90*, 1872-1876.

Miller,J.M. (2000). Craniofacial variation in Homo habilis: an analysis of the evidence for multiple species. Am. J. Phys. Anthropol. *112*, 103-128.

Mondal,H. and Hofschneider,P.H. (1982). Isolation and characterization of retrovirus-like elements from normal human fetuses. Int. J. Cancer *30*, 281-287.

Moya-Sola,S. and Kohler,M. (1996). A Dryopithecus skeleton and the origins of great-ape locomotion. Nature *379*, 156-159.

Mueller-Lantzsch,N., Sauter,M., Weiskircher,A., Kramer,K., Best,B., Buck,M., and Grasser,F. (1993). Human endogenous retroviral element K10 (HERV-K10) encodes a full-length gag homologous 73-kDa protein and a functional protease. AIDS Res. Hum. Retroviruses *9*, 343-350.

Myers,J.S., Vincent,B.J., Udall,H., Watkins,W.S., Morrish,T.A., Kilroy,G.E., Swergold,G.D., Henke,J., Henke,L., Moran,J.V., Jorde,L.B., and Batzer,M.A. (2002). A comprehensive analysis of recently integrated human Ta L1 elements Am. J. Hum. Genet. *71*, 312-326.

Navarro,A. and Barton,N.H. (2003). Chromosomal speciation and molecular divergence--accelerated evolution in rearranged chromosomes. Science *300*, 321-324.

Nei,M. (1973). Analysis of gene diversity in subdivided populations. Proc. Natl. Acad. Sci. U. S. A *70*, 3321-3323.

Nei,M. (1978). Estimation of average heterozygosity and genetic distance from a small number of individuals. Genetics *89*, 583-590.

Newman,T. and Trask,B.J. (2003). Complex evolution of 7E olfactory receptor genes in segmental duplications. Genome Res. *13*, 781-793.

Nickerson,D.A., Taylor,S.L., Weiss,K.M., Clark,A.G., Hutchinson,R.G., Stengard,J., Salomaa,V., Vartiainen,E., Boerwinkle,E., and Sing,C.F. (1998). DNA sequence diversity in a 9.7-kb region of the human lipoprotein lipase gene. Nat. Genet. *19*, 233-240.

Noda,R., Kim,C.G., Takenaka,O., Ferrell,R.E., Tanoue,T., Hayasaka,I., Ueda,S., Ishida,T., and Saitou,N. (2001). Mitochondrial 16S rRNA sequence diversity of hominoids. J. Hered. *92*, 490-496.

Nordborg,M. (1998). On the probability of Neanderthal ancestry. Am. J. Hum. Genet. *63*, 1237-1240.

Ono,M. (1986). Molecular cloning and long terminal repeat sequences of human endogenous retrovirus genes related to types A and B retrovirus genes. J. Virol. *58*, 937-944.

Ono,M., Kawakami,M., and Takezawa,T. (1987). A novel human nonviral retroposon derived from an endogenous retrovirus. Nucleic Acids Res. *15*, 8725-8737.

Ono,M., Yasunaga,T., Miyata,T., and Ushikubo,H. (1986). Nucleotide sequence of human endogenous retrovirus genome related to the mouse mammary tumor virus genome. J. Virol. *60*, 589-598.

Ostertag,E.M., Goodier,J.L., Zhang,Y., and Kazazian,H.H., Jr. (2003). SVA elements are nonautonomous retrotransposons that cause disease in humans. Am. J. Hum. Genet. *73*, 1444-1451.

Ovchinnikov,I., Rubin,A., and Swergold,G.D. (2002). Tracing the LINEs of human evolution. Proc. Natl. Acad. Sci. U. S. A *99*, 10522-10527.

Ovchinnikov,I.V., Gotherstrom,A., Romanova,G.P., Kharitonov,V.M., Liden,K., and Goodwin,W. (2000). Molecular analysis of Neanderthal DNA from the northern Caucasus. Nature *404*, 490-493.

Patience,C., Simpson,G.R., Colletta,A.A., Welch,H.M., Weiss,R.A., and Boyd,M.T. (1996). Human endogenous retrovirus expression and reverse transcriptase activity in the T47D mammary carcinoma cell line. J. Virol. *70*, 2654-2657.

Patience,C., Wilkinson,D.A., and Weiss,R.A. (1997). Our retroviral heritage. Trends Genet. *13*, 116-120.

Pavelitz,T., Liao,D., and Weiner,A.M. (1999). Concerted evolution of the tandem array encoding primate U2 snRNA (the RNU2 locus) is accompanied by dramatic remodeling of the junctions with flanking chromosomal sequences. EMBO J. *18*, 3783-3792.

Pavlicek,A., Paces,J., Elleder,D., and Hejnar,J. (2002). Processed pseudogenes of human endogenous retroviruses generated by LINEs: their integration, stability, and distribution. Genome Res. *12*, 391-399.

Peakall, R. and Smouse, P. E. GenAIX V5: Genetic Analysis in Excel. Population genetic software for teaching and research. Australian National University, Canberra, Australia. 2001. (Computer Program)

Perez-Perez,A., Bermudez de Castro,J.M., and Arsuaga,J.L. (1999). Nonocclusal dental microwear analysis of 300,000-year-old Homo heilderbergensis teeth from Sima de los Huesos (Sierra de Atapuerca, Spain). Am. J. Phys. Anthropol. *108*, 433-457.

Pickford,M. (2001). Discovery of earliest hominid remains. Science *291*, 986.

Pitulko,V.V., Nikolsky,P.A., Girya,E.Y., Basilyan,A.E., Tumskoy,V.E., Koulakov,S.A., Astakhov,S.N., Pavlova,E.Y., and Anisimov,M.A. (2004). The Yana RHS site: humans in the Arctic before the last glacial maximum. Science *303*, 52-56.

Pope,K.O., D'Hondt,S.L., and Marshall,C.R. (1998). Meteorite impact and the mass extinction of species at the Cretaceous/Tertiary boundary. Proc. Natl. Acad. Sci. U. S. A *95*, 11028-11029.

Redd,A.J. and Stoneking,M. (1999). Peopling of Sahul: mtDNA variation in aboriginal Australian and Papua New Guinean populations. Am. J. Hum. Genet. *65*, 808-828.

Relethford,J.H. and Harpending,H.C. (1994). Craniometric variation, genetic theory, and modern human origins. Am. J. Phys. Anthropol. *95*, 249-270.

Relethford,J.H. and Jorde,L.B. (1999). Genetic evidence for larger African population size during recent human evolution. Am. J. Phys. Anthropol. *108*, 251-260.

Reus,K., Mayer,J., Sauter,M., Scherer,D., Muller-Lantzsch,N., and Meese,E. (2001a). Genomic organization of the human endogenous retrovirus HERV-K(HML-2.HOM) (ERVK6) on chromosome 7. Genomics *72*, 314-320.

Reus,K., Mayer,J., Sauter,M., Zischler,H., Muller-Lantzsch,N., and Meese,E. (2001b). HERV-K(OLD): ancestor sequences of the human endogenous retrovirus family HERV-K(HML-2). J. Virol. *75*, 8917-8926.

Roberts-Thomson,J.M., Martinson,J.J., Norwich,J.T., Harding,R.M., Clegg,J.B., and Boettcher,B. (1996). An ancient common origin of aboriginal Australians and New Guinea highlanders is supported by alpha-globin haplotype analysis. Am. J. Hum. Genet. *58*, 1017-1024.

Roberts,M.B., Stringer,C.B., and Parfitt,S.A. (1994). A hominid tibia from Middle Pleistocene sediments at Boxgrove, UK. Nature *369*, 311-313.

Roy-Engel,A.M., Carroll,M.L., El Sawy,M., Salem,A.H., Garber,R.K., Nguyen,S.V., Deininger,P.L., and Batzer,M.A. (2002). Non-traditional Alu evolution and primate genomic diversity. J. Mol. Biol. *316*, 1033-1040.

Roy-Engel,A.M., Carroll,M.L., Vogel,E., Garber,R.K., Nguyen,S.V., Salem,A.H., Batzer,M.A., and Deininger,P.L. (2001). Alu insertion polymorphisms for the study of human genomic diversity. Genetics *159*, 279-290.

Ruda,V.M., Akopov,S.B., Trubetskoy,D.O., Manuylov,N.L., Vetchinova,A.S., Zavalova,L.L., Nikolaev,L.G., and Sverdlov,E.D. (2004). Tissue specificity of enhancer and promoter activities of a HERV-K(HML-2) LTR. Virus Res. *104*, 11-16.

Saitou,N. and Nei,M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. *4*, 406-425.

Salem,A.H., Kilroy,G.E., Watkins,W.S., Jorde,L.B., and Batzer,M.A. (2003a). Recently integrated Alu elements and human genomic diversity. Mol. Biol. Evol. *20*, 1349-1361.

Salem,A.H., Myers,J.S., Otieno,A.C., Watkins,W.S., Jorde,L.B., and Batzer,M.A. (2003b). LINE-1 preTa elements in the human genome. J. Mol. Biol. *326*, 1127-1146.

Salem,A.H., Ray,D.A., Xing,J., Callinan,P.A., Myers,J.S., Hedges,D.J., Garber,R.K., Witherspoon,D.J., Jorde,L.B., and Batzer,M.A. (2003c). Alu elements and hominid phylogenetics. Proc. Natl. Acad. Sci. U. S. A *100*, 12787-12791.

Sanger,F., Nicklen,S., and Coulson,A.R. (1977). DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. U. S. A *74*, 5463-5467.

Sarich,V.M. and Wilson,A.C. (1967). Rates of albumin evolution in primates. Proc. Natl. Acad. Sci. U. S. A *58*, 142-148.

Sassaman,D.M., Dombroski,B.A., Moran,J.V., Kimberland,M.L., Naas,T.P., DeBerardinis,R.J., Gabriel,A., Swergold,G.D., and Kazazian,H.H., Jr. (1997). Many human L1 elements are capable of retrotransposition. Nat. Genet. *16*, 37-43.

Sauter,M., Schommer,S., Kremmer,E., Remberger,K., Dolken,G., Lemm,I., Buck,M., Best,B., Neumann-Haefelin,D., and Mueller-Lantzsch,N. (1995). Human

endogenous retrovirus K10: expression of Gag protein and detection of antibodies in patients with seminomas. J. Virol. *69*, 414-421.

Schmitz,J., Ohme,M., and Zischler,H. (2001). SINE insertions in cladistic analyses and the phylogenetic affiliations of Tarsius bancanus to other primates. Genetics *157*, 777-784.

Schmitz,R.W., Serre,D., Bonani,G., Feine,S., Hillgruber,F., Krainitzki,H., Paabo,S., and Smith,F.H. (2002). The Neandertal type site revisited: interdisciplinary investigations of skeletal remains from the Neander Valley, Germany. Proc. Natl. Acad. Sci. U. S. A *99*, 13342-13347.

Schommer,S., Sauter,M., Krausslich,H.G., Best,B., and Mueller-Lantzsch,N. (1996). Characterization of the human endogenous retrovirus K proteinase. J. Gen. Virol. *77*, 375-379.

Schrago,C.G. and Russo,C.A. (2003). Timing the origin of new world monkeys. Mol. Biol. Evol. *20*, 1620-1625.

Schwartz,A., Chan,D.C., Brown,L.G., Alagappan,R., Pettay,D., Disteche,C., McGillivray,B., de la,C.A., and Page,D.C. (1998). Reconstructing hominid Y evolution: X-homologous block, created by X-Y transposition, was disrupted by Yp inversion through LINE-LINE recombination. Hum. Mol. Genet. *7*, 1-11.

Seifarth,W., Baust,C., Murr,A., Skladny,H., Krieg-Schneider,F., Blusch,J., Werner,T., Hehlmann,R., and Leib-Mosch,C. (1998). Proviral structure, chromosomal location, and expression of HERV-K- T47D, a novel human endogenous retrovirus derived from T47D particles J. Virol. *72*, 8384-8391.

Seifarth,W., Skladny,H., Krieg-Schneider,F., Reichert,A., Hehlmann,R., and Leib-Mosch,C. (1995). Retrovirus-like particles released from the human breast cancer cell line T47-D display type B- and C-related endogenous retroviral sequences. J. Virol. *69*, 6408-6416.

Seiffert,E.R., Simons,E.L., and Attia,Y. (2003). Fossil evidence for an ancient divergence of lorises and galagos. Nature *422*, 421-424.

Semino,O., Passarino,G., Oefner,P.J., Lin,A.A., Arbuzova,S., Beckman,L.E., De Benedictis,G., Francalacci,P., Kouvatsi,A., Limborska,S., Marcikiae,M., Mika,A., Mika,B., Primorac,D., Santachiara-Benerecetti,A.S., Cavalli-Sforza,L.L., and Underhill,P.A. (2000). The genetic legacy of Paleolithic Homo sapiens sapiens in extant Europeans: a Y chromosome perspective. Science *290*, 1155-1159.

Serre,D., Langaney,A., Chech,M., Teschler-Nicola,M., Paunovic,M., Mennecier,P., Hofreiter,M., Possnert,G.G., and Paabo,S. (2004). No Evidence of Neandertal mtDNA Contribution to Early Modern Humans. PLoS. Biol. *2*, E57.

Sheen,F.M., Sherry,S.T., Risch,G.M., Robichaux,M., Nasidze,I., Stoneking,M., Batzer,M.A., and Swergold,G.D. (2000). Reading between the LINEs: human genomic variation induced by LINE-1 retrotransposition. Genome Res. *10*, 1496-1508.

Shoshani,J., Groves,C.P., Simons,E.L., and Gunnell,G.F. (1996). Primate phylogeny: morphological vs. molecular results. Mol. Phylogenet. Evol. *5*, 102-154.

Sibley,C.G. and Ahlquist,J.E. (1984). The phylogeny of the hominoid primates, as indicated by DNA-DNA hybridization. J. Mol. Evol. *20*, 2-15.

Sibley,C.G. and Ahlquist,J.E. (1987). DNA hybridization evidence of hominoid phylogeny: results from an expanded data set. J. Mol. Evol. *26*, 99-121.

Simmonds,P., Tuplin,A., and Evans,D.J. (2004). Detection of genome-scale ordered RNA structure (GORS) in genomes of positive-stranded RNA viruses: Implications for virus evolution and host persistence. RNA. *10*, 1337-1351.

Simpson,G.R., Patience,C., Lower,R., Tonjes,R.R., Moore,H.D., Weiss,R.A., and Boyd,M.T. (1996). Endogenous D-type (HERV-K) related sequences are packaged into retroviral particles in the placenta and possess open reading frames for reverse transcriptase. Virology *222*, 451-456.

Smith,C.I., Chamberlain,A.T., Riley,M.S., Stringer,C., and Collins,M.J. (2003). The thermal history of human fossils and the likelihood of successful DNA amplification. J. Hum. Evol. *45*, 203-217.

Smith,F.H. (1985). Continuity and change in the origin of modern Homo sapiens. Z. Morphol. Anthropol. *75*, 197-222.

Smith,F.H., Boyd,D.C., and Malez,M. (1985). Additional upper Pleistocene human remains from Vindija cave, Croatia, Yugoslavia. Am. J. Phys. Anthropol. *68*, 375-383.

Stankiewicz,P. and Lupski,J.R. (2002a). Genome architecture, rearrangements and genomic disorders. Trends Genet. *18*, 74-82.

Stankiewicz,P. and Lupski,J.R. (2002b). Molecular-evolutionary mechanisms for genomic disorders. Curr. Opin. Genet. Dev. *12*, 312-319.

Stauffer,R.L., Walker,A., Ryder,O.A., Lyons-Weiler,M., and Hedges,S.B. (2001). Human and ape molecular clocks and constraints on paleontological hypotheses. J. Hered. *92*, 469-474.

Steinhuber,S., Brack,M., Hunsmann,G., Schwelberger,H., Dierich,M.P., and Vogetseder,W. (1995). Distribution of human endogenous retrovirus HERV-K genomes in humans and different primates. Hum. Genet. *96*, 188-192.

Stoneking,M., Fontius,J.J., Clifford,S.L., Soodyall,H., Arcot,S.S., Saha,N., Jenkins,T., Tahir,M.A., Deininger,P.L., and Batzer,M.A. (1997). Alu insertion polymorphisms and human evolution: evidence for a larger population size in Africa Genome Res. *7*, 1061-1071.

Stringer,C. (2003). Human evolution: Out of Ethiopia. Nature *423*, 692-3, 695.

Sugimoto,J., Matsuura,N., Kinjo,Y., Takasu,N., Oda,T., and Jinno,Y. (2001). Transcriptionally active HERV-K genes: identification, isolation, and chromosomal mapping. Genomics *72*, 137-144.

Sun,C., Skaletsky,H., Rozen,S., Gromoll,J., Nieschlag,E., Oates,R., and Page,D.C. (2000). Deletion of azoospermia factor a (AZFa) region of human Y chromosome caused by recombination between HERV15 proviruses. Hum. Mol. Genet. *9*, 2291-2296.

Sverdlov,E. (2005). Retroviruses and Primate Genome Evolution. Landes Bioscience).

Sverdlov,E.D. (2000). Retroviruses and primate evolution. Bioessays *22*, 161-171.

Swisher,C.C., III, Rink,W.J., Anton,S.C., Schwarcz,H.P., Curtis,G.H., Suprijo,A., and Widiasmoro (1996). Latest Homo erectus of Java: potential contemporaneity with Homo sapiens in southeast Asia. Science *274*, 1870-1874.

Takahata,N. and Satta,Y. (1997). Evolution of the primate lineage leading to modern humans: phylogenetic and demographic inferences from DNA sequences. Proc. Natl. Acad. Sci. U. S. A *94*, 4811-4815.

Takai,M., Anaya,F., Shigehara,N., and Setoguchi,T. (2000). New fossil materials of the earliest new world monkey, Branisella boliviana, and the problem of platyrrhine origins. Am. J. Phys. Anthropol. *111*, 263-281.

Tassabehji,M., Strachan,T., Anderson,M., Campbell,R.D., Collier,S., and Lako,M. (1994). Identification of a novel family of human endogenous retroviruses and characterization of one family member, HERV-K(C4), located in the complement C4 gene cluster. Nucleic Acids Res. *22*, 5211-5217.

Templeton,A. (2002). Out of Africa again and again. Nature *416*, 45-51.

Tishkoff,S.A., Dietzsch,E., Speed,W., Pakstis,A.J., Kidd,J.R., Cheung,K., Bonne-Tamir,B., Santachiara-Benerecetti,A.S., Moral,P., and Krings,M. (1996). Global patterns of linkage disequilibrium at the CD4 locus and modern human origins. Science *271*, 1380-1387.

Tonjes,R.R., Boller,K., Limbach,C., Lugert,R., and Kurth,R. (1997). Characterization of human endogenous retrovirus type K virus-like particles generated from recombinant baculoviruses. Virology *233*, 280-291.

Tonjes,R.R., Czauderna,F., and Kurth,R. (1999). Genome-wide screening, cloning, chromosomal assignment, and expression of full-length human endogenous retrovirus type K. J. Virol. *73*, 9187-9195.

Towler,E.M., Gulnik,S.V., Bhat,T.N., Xie,D., Gustschina,E., Sumpter,T.R., Robertson,N., Jones,C., Sauter,M., Mueller-Lantzsch,N., Debouck,C., and Erickson,J.W. (1998). Functional characterization of the protease of human endogenous retrovirus, K10: can it complement HIV-1 protease? Biochemistry *37*, 17137-17144.

Trask,B.J., Friedman,C., Martin-Gallardo,A., Rowen,L., Akinbami,C., Blankenship,J., Collins,C., Giorgi,D., Iadonato,S., Johnson,F., Kuo,W.L., Massa,H., Morrish,T., Naylor,S., Nguyen,O.T., Rouquier,S., Smith,T., Wong,D.J., Youngblom,J., and van den,E.G. (1998). Members of the olfactory receptor gene family are contained in large blocks of DNA duplicated polymorphically near the ends of human chromosomes. Hum. Mol. Genet. *7*, 13-26.

Tremblay,A., Jasin,M., and Chartrand,P. (2000). A double-strand break in a chromosomal LINE element can be repaired by gene conversion with various endogenous LINE elements in mouse cells. Mol. Cell Biol. *20*, 54-60.

Trinkaus,E., Moldovan,O., Milota,S., Bilgar,A., Sarcina,L., Athreya,S., Bailey,S.E., Rodrigo,R., Mircea,G., Higham,T., Ramsey,C.B., and van der,P.J. (2003). An early modern human from the Pestera cu Oase, Romania. Proc. Natl. Acad. Sci. U. S. A *100*, 11231-11236.

Tristem,M. (2000). Identification and characterization of novel human endogenous retrovirus families by phylogenetic screening of the human genome mapping project database. J. Virol. *74*, 3715-3730.

Tsend-Ayush,E., Grutzner,F., Yue,Y., Grossmann,B., Hansel,U., Sudbrak,R., and Haaf,T. (2004). Plasticity of human chromosome 3 during primate evolution. Genomics *83*, 193-202.

Turner,G., Barbulescu,M., Su,M., Jensen-Seaman,M.I., Kidd,K.K., and Lenz,J. (2001). Insertional polymorphisms of full-length endogenous retroviruses in humans. Curr. Biol. *11*, 1531-1535.

Van der Kuyl,A.C., Kuiken,C.L., Dekker,J.T., and Goudsmit,J. (1995). Phylogeny of African monkeys based upon mitochondrial 12S rRNA sequences. J. Mol. Evol. *40*, 173-180.

Vekua,A., Lordkipanidze,D., Rightmire,G.P., Agusti,J., Ferring,R., Maisuradze,G., Mouskhelishvili,A., Nioradze,M., De Leon,M.P., Tappen,M., Tvalchrelidze,M., and Zollikofer,C. (2002). A new skull of early Homo from Dmanisi, Georgia. Science *297*, 85-89.

Vignaud,P., Duringer,P., Mackaye,H.T., Likius,A., Blondel,C., Boisserie,J.R., de Bonis,L., Eisenmann,V., Etienne,M.E., Geraads,D., Guy,F., Lehmann,T., Lihoreau,F., Lopez-Martinez,N., Mourer-Chauvire,C., Otero,O., Rage,J.C., Schuster,M., Viriot,L., Zazzo,A., and Brunet,M. (2002). Geology and palaeontology of the Upper Miocene Toros-Menalla hominid locality, Chad. Nature *418*, 152-155.

Vincent,B.J., Myers,J.S., Ho,H.J., Kilroy,G.E., Walker,J.A., Watkins,W.S., Jorde,L.B., and Batzer,M.A. (2003). Following the LINEs: an analysis of primate genomic variation at human-specific LINE-1 insertion sites. Mol. Biol. Evol. *20*, 1338-1348.

Vinogradova,T.V., Leppik,L.P., Nikolaev,L.G., Akopov,S.B., Kleiman,A.M., Senyuta,N.B., and Sverdlov,E.D. (2001). Solitary human endogenous retroviruses-K LTRs retain transcriptional activity in vivo, the mode of which is different in different cell types Virology *290*, 83-90.

Voisset, C. and Andrawiss, M. Retroviruses at a glance. 1(3), 4015.1-4015.4. 2000. Genome Biology.

Voisset,C., Blancher,A., Perron,H., Mandrand,B., Mallet,F., and Paranhos-Baccala,G. (1999). Phylogeny of a novel family of human endogenous retrovirus sequences, HERV-W, in humans and other primates. AIDS Res. Hum. Retroviruses *15*, 1529-1533.

Waldman,A.S. and Liskay,R.M. (1988). Dependence of intrachromosomal recombination in mammalian cells on uninterrupted homology. Mol. Cell Biol. *8*, 5350-5357.

Wall,J.D. (2000). Detecting ancient admixture in humans using sequence polymorphism data. Genetics *154*, 1271-1279.

Walter,R.C., Buffler,R.T., Bruggemann,J.H., Guillaume,M.M., Berhe,S.M., Negassi,B., Libsekal,Y., Cheng,H., Edwards,R.L., von Cosel,R., Neraudeau,D., and Gagnon,M. (2000). Early human occupation of the Red Sea coast of Eritrea during the last interglacial. Nature *405*, 65-69.

Wang-Johanning,F., Frost,A.R., Jian,B., Epp,L., Lu,D.W., and Johanning,G.L. (2003). Quantitation of HERV-K env gene expression and splicing in human breast cancer. Oncogene *22*, 1528-1535.

Wang-Johanning,F., Frost,A.R., Johanning,G.L., Khazaeli,M.B., LoBuglio,A.F., Shaw,D.R., and Strong,T.V. (2001). Expression of human endogenous retrovirus k envelope transcripts in human breast cancer. Clin. Cancer Res. *7*, 1553-1560.

Watkins,W.S., Rogers,A.R., Ostler,C.T., Wooding,S., Bamshad,M.J., Brassington,A.M., Carroll,M.L., Nguyen,S.V., Walker,J.A., Prasad,B.V., Reddy,P.G., Das,P.K., Batzer,M.A., and Jorde,L.B. (2003). Genetic variation among world populations: inferences from 100 Alu insertion polymorphisms. Genome Res. *13*, 1607-1618.

Wei,W., Gilbert,N., Ooi,S.L., Lawler,J.F., Ostertag,E.M., Kazazian,H.H., Boeke,J.D., and Moran,J.V. (2001). Human L1 retrotransposition: cis preference versus trans complementation. Mol. Cell Biol. *21*, 1429-1439.

White,T.D., Asfaw,B., DeGusta,D., Gilbert,H., Richards,G.D., Suwa,G., and Howell,F.C. (2003). Pleistocene Homo sapiens from Middle Awash, Ethiopia. Nature *423*, 742-747.

White,T.D., Suwa,G., and Asfaw,B. (1994). Australopithecus ramidus, a new species of early hominid from Aramis, Ethiopia. Nature *371*, 306-312.

Wolpoff,M.H. (1996). Interpretations of multiregional evolution. Science *274*, 704-707.

Wolpoff,M.H., Hawks,J., and Caspari,R. (2000). Multiregional, not multiple origins. Am. J. Phys. Anthropol. *112*, 129-136.

Wood,B. (2002). Hominid revelations from Chad. Nature *418*, 133-135.

Yeh, F. C, Yang, R. C, Boyle, T., Ye, Z. H, and Mao, J. X. PopGene, the user-friendly shareware for population genetic analysis. Molecular Biology and Biotechnology Centre, University of Alberta, Canada. 1997. (Computer Program)

Yi,J.M., Takenaka,O., and Kim,H.S. (2003). Molecular characterization and phylogenetic relationship of HERV-W family in the Macaca fuscata. Arch. Virol. *148*, 1613-1622.

Yoder,A.D. and Yang,Z. (2000). Estimation of primate speciation dates using local molecular clocks. Mol. Biol. Evol. *17*, 1081-1090.

Young,N.M. and MacLatchy,L. (2004). The phylogenetic position of Morotopithecus. J. Hum. Evol. *46*, 163-184.

Yu,H., Jetzt,A.E., Ron,Y., Preston,B.D., and Dougherty,J.P. (1998). The nature of human immunodeficiency virus type 1 strand transfers. J. Biol. Chem. *273*, 28384-28391.

Yu,N., Chen,F.C., Ota,S., Jorde,L.B., Pamilo,P., Patthy,L., Ramsay,M., Jenkins,T., Shyue,S.K., and Li,W.H. (2002). Larger genetic differences within africans than between Africans and Eurasians. Genetics *161*, 269-274.

Yu,N., Zhao,Z., Fu,Y.X., Sambuughin,N., Ramsay,M., Jenkins,T., Leskinen,E., Patthy,L., Jorde,L.B., Kuromori,T., and Li,W.H. (2001). Global patterns of human DNA sequence variation in a 10-kb region on chromosome 1. Mol. Biol. Evol. *18*, 214-222.

Yunis,J.J. and Prakash,O. (1982). The origin of man: a chromosomal pictorial legacy. Science. *19;215*, 1525-1530.

Zhao,Z., Jin,L., Fu,Y.X., Ramsay,M., Jenkins,T., Leskinen,E., Pamilo,P., Trexler,M., Patthy,L., Jorde,L.B., Ramos-Onsins,S., Yu,N., and Li,W.H. (2000). Worldwide DNA sequence variation in a 10-kilobase noncoding region on human chromosome 22. Proc. Natl. Acad. Sci. U. S. A *97*, 11354-11358.

Zhu,Z.B., Jian,B., and Volanakis,J.E. (1994). Ancestry of SINE-R.C2 a human-specific retroposon. Hum. Genet. *93*, 545-551.

Zietkiewicz,E., Yotova,V., Jarnik,M., Korab-Laskowska,M., Kidd,K.K., Modiano,D., Scozzari,R., Stoneking,M., Tishkoff,S., Batzer,M., and Labuda,D. (1997). Nuclear DNA diversity in worldwide distributed human populations. Gene *205*, 161-171.

Zsiros,J., Jebbink,M.F., Lukashov,V.V., Voute,P.A., and Berkhout,B. (1998). Evolutionary relationships within a subgroup of HERV-K-related human endogenous retroviruses. J. Gen. Virol. *79 (Pt 1)*, 61-70.

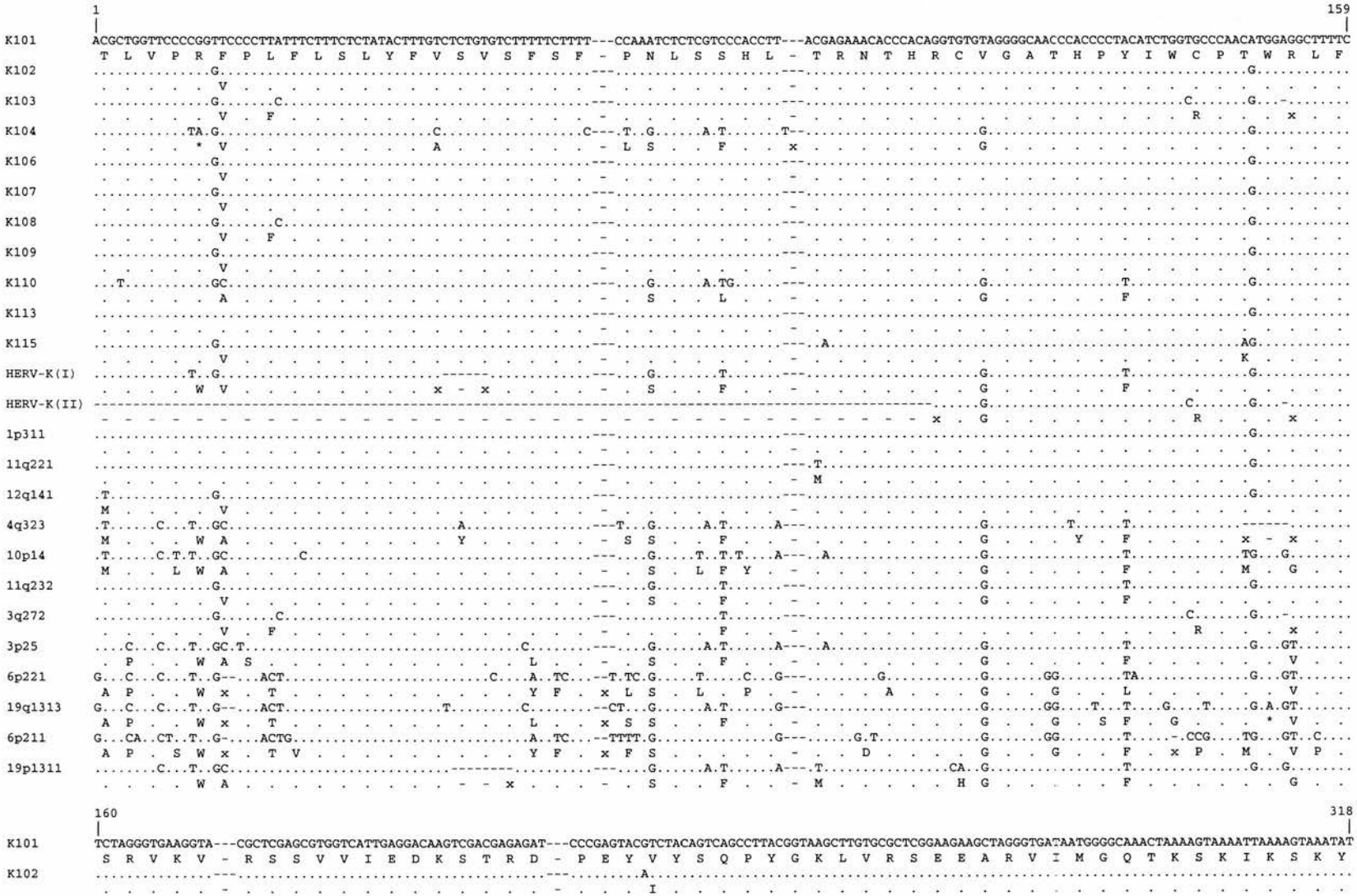
Zsiros,J., Jebbink,M.F., Lukashov,V.V., Voute,P.A., and Berkhout,B. (1999). Biased nucleotide composition of the genome of HERV-K related endogenous retroviruses and its evolutionary implications. J. Mol. Evol. *48*, 102-111.

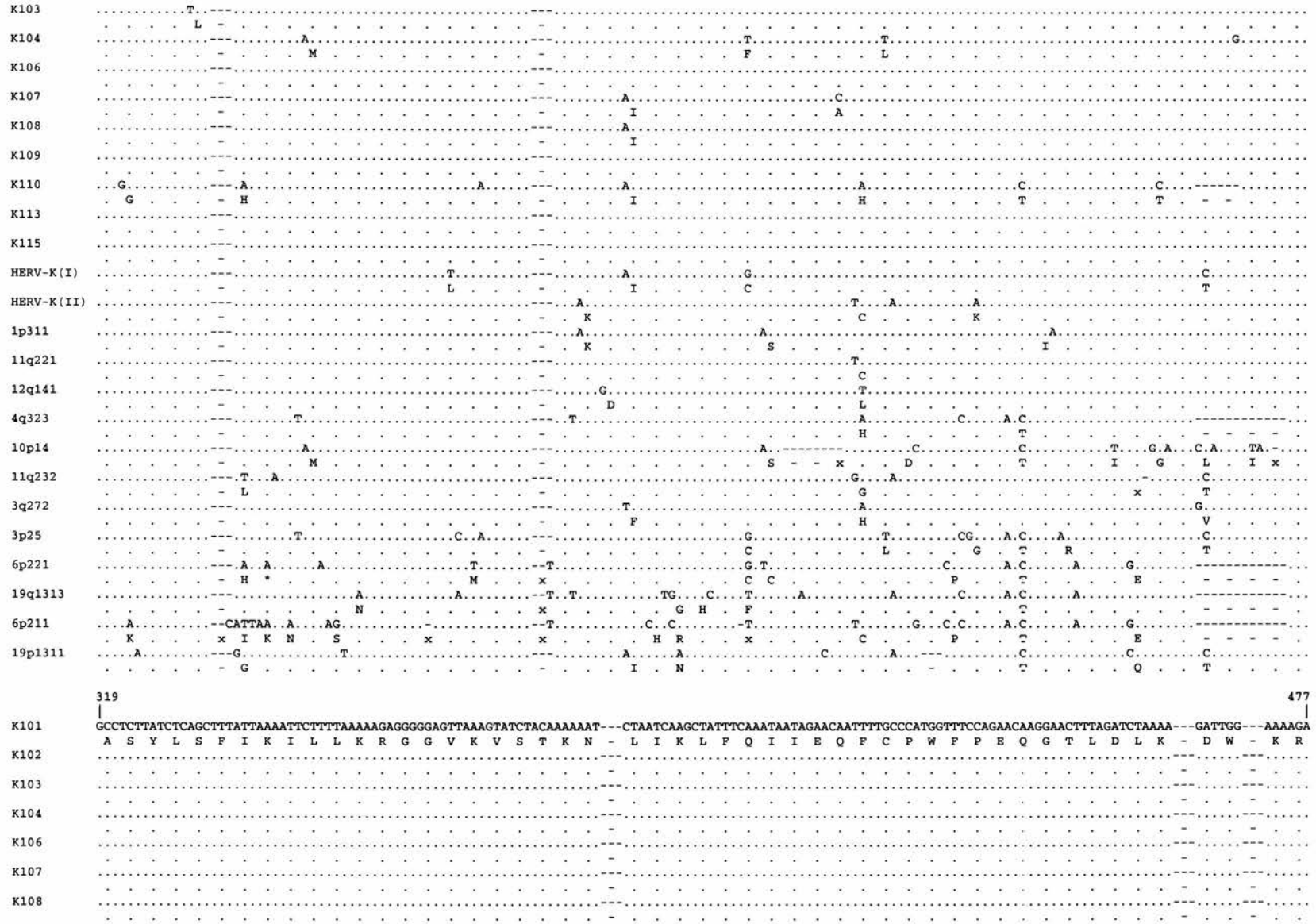
APPENDIX A

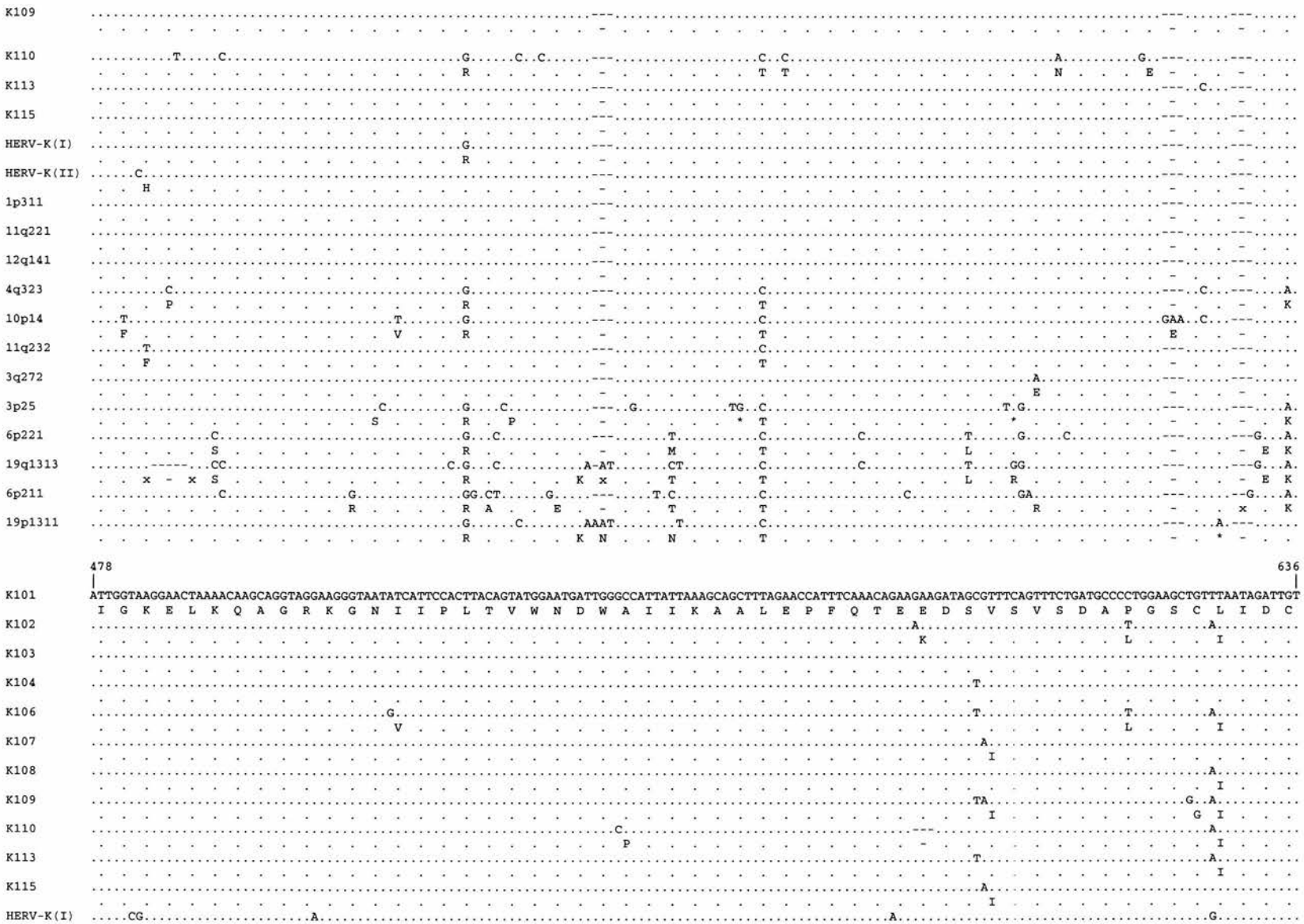
SEQUENCE ALIGMENTS

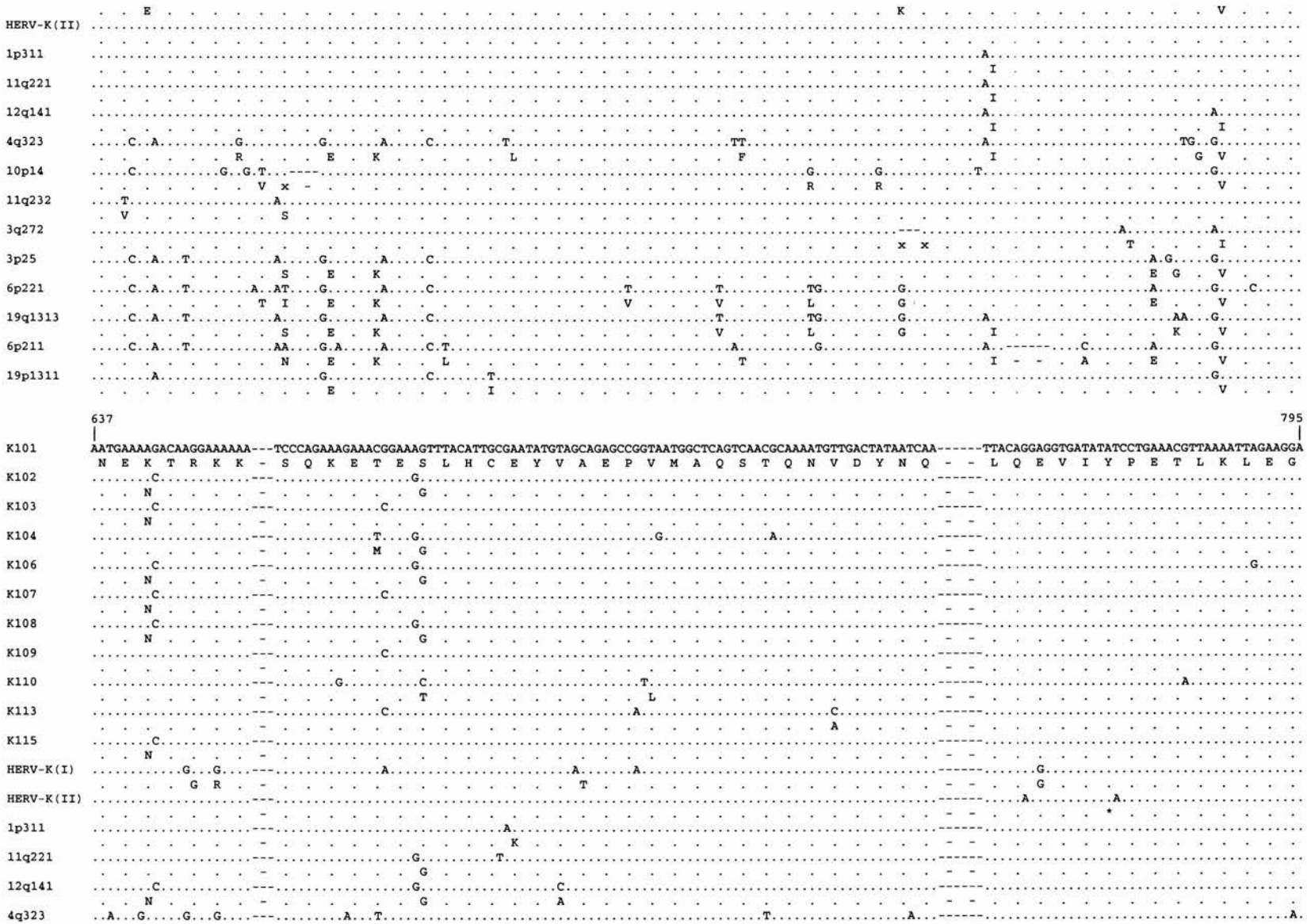
Table of Contents

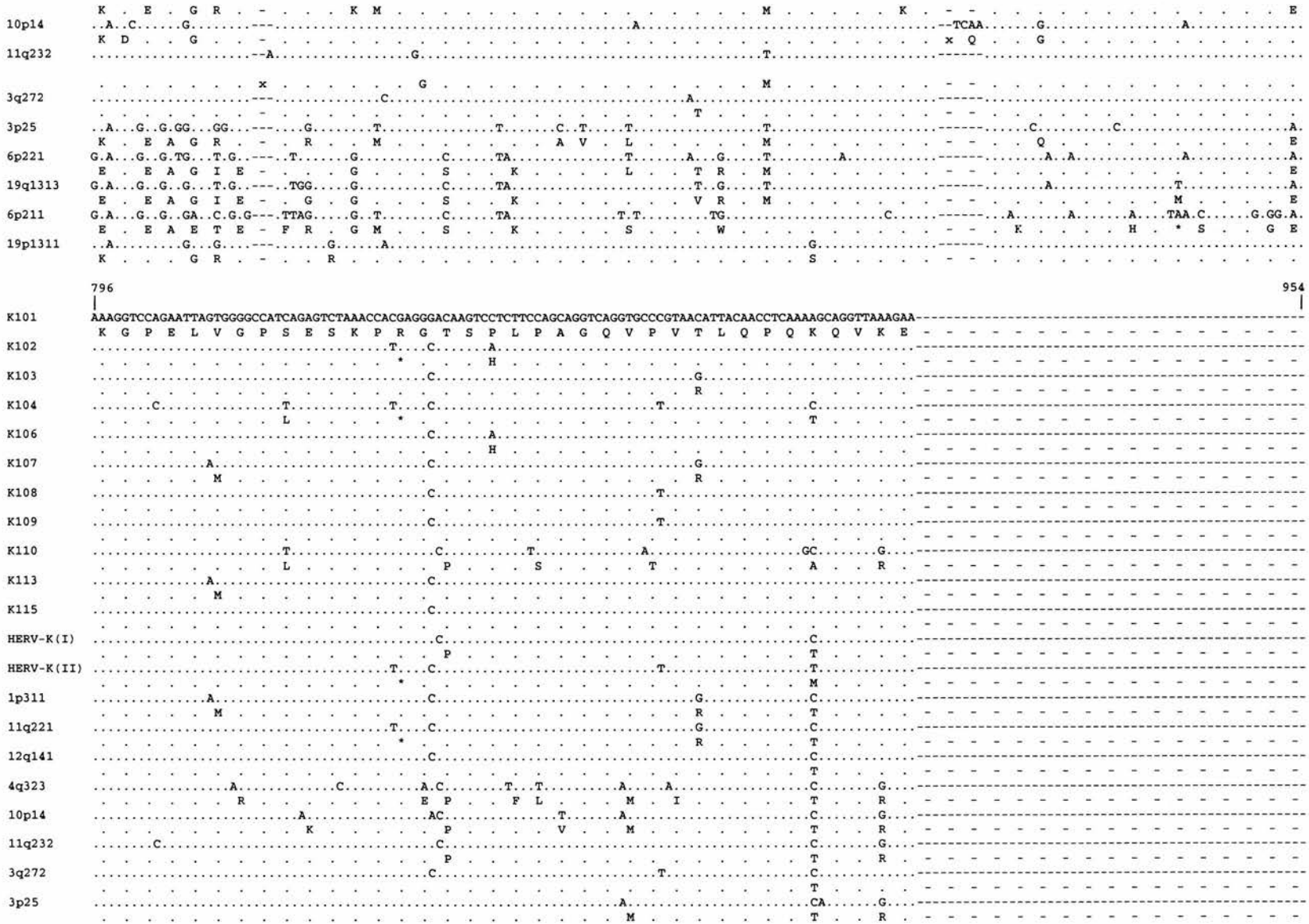
A.1 HERV-K(HML-2) Alignment of putative proviral ORFs

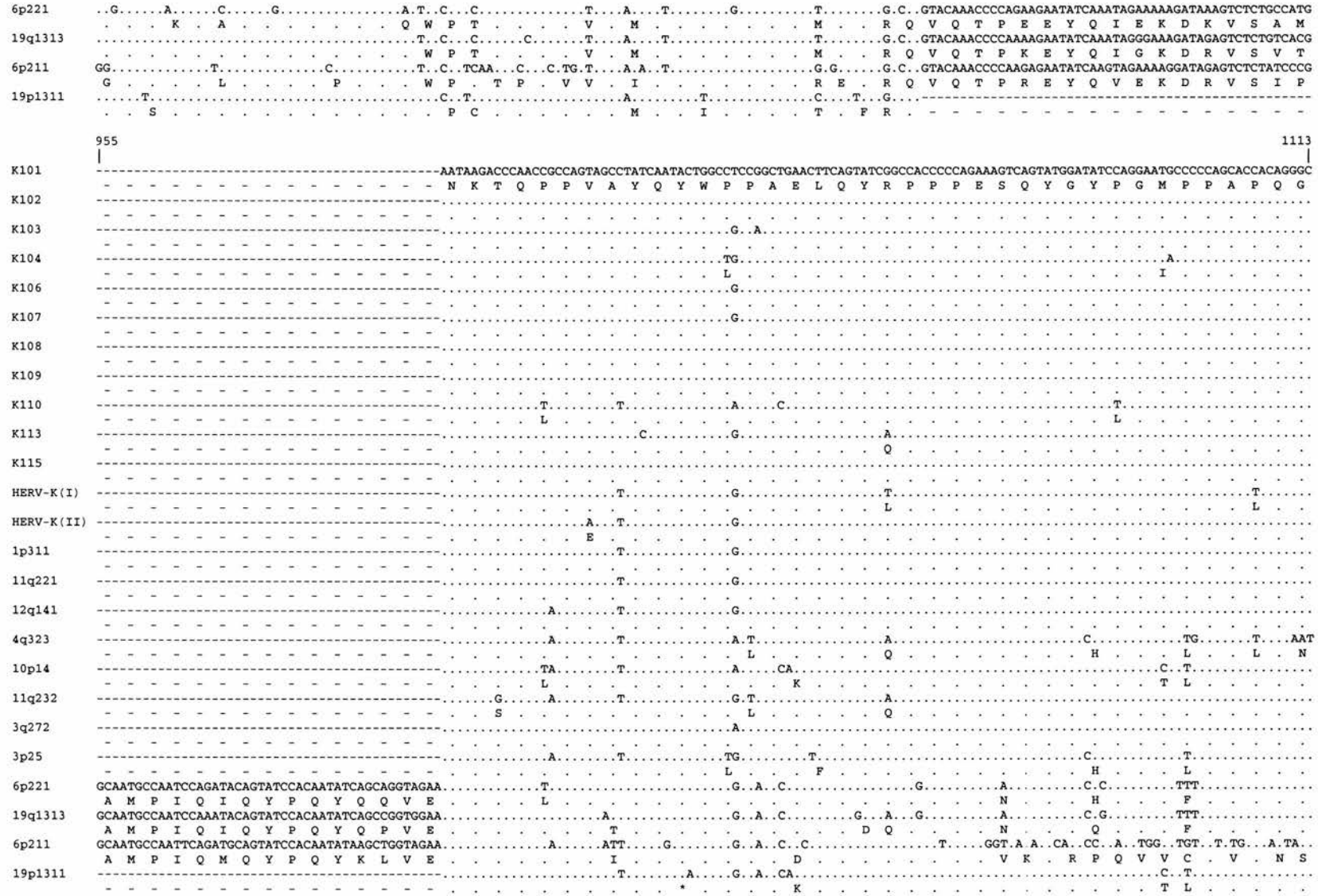


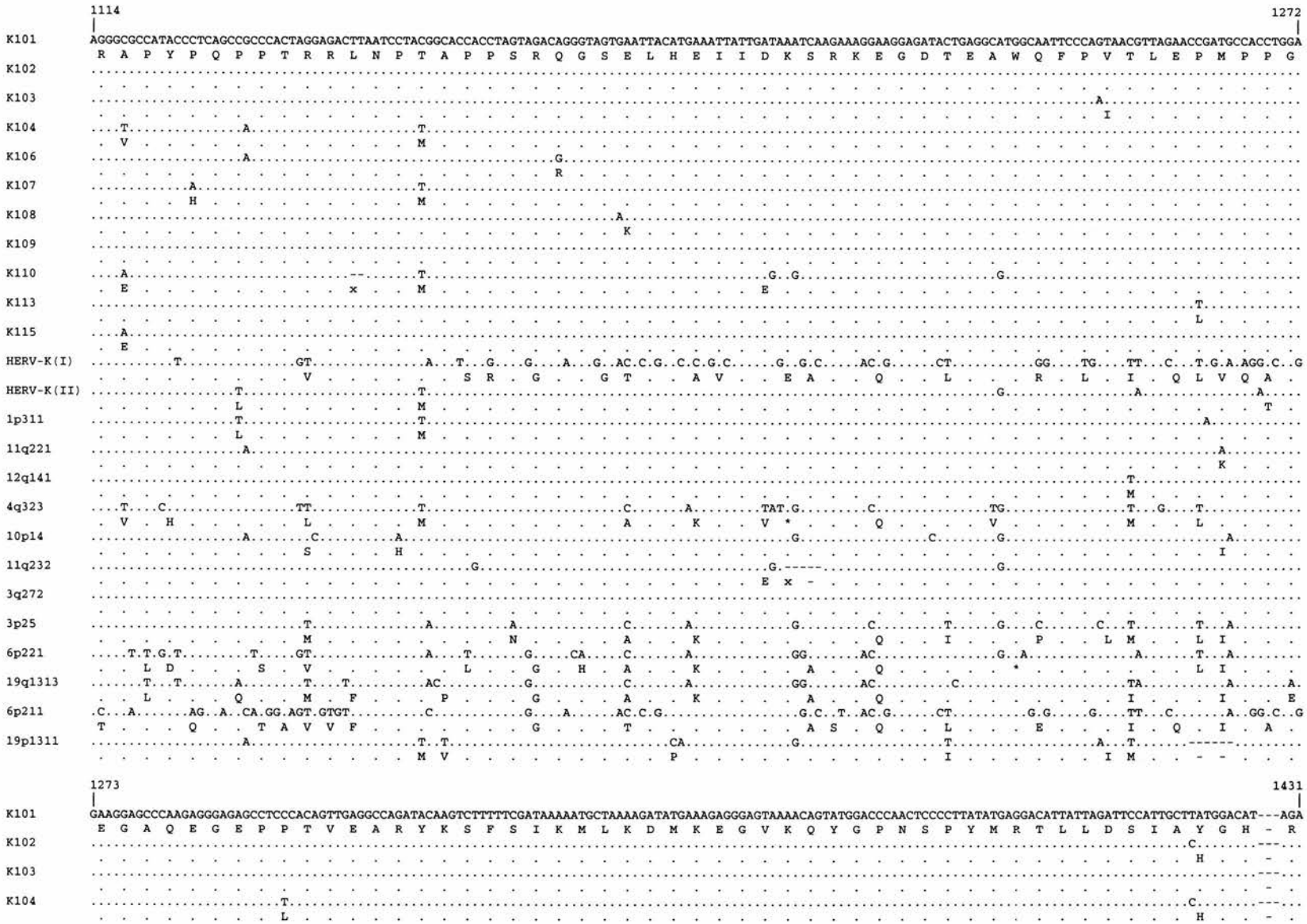

K103
K104
K106
K107
K108
K109
K110
K113
K115
HERV-K(I)
HERV-K(II)
1p311
11q221
12q141
4q323
10p14
11q232
3q272
3p25
6p221
19q1313
6p211
19p1311
319
477
K101
GCCTCTTATCTCAGCTTTATTAAAATTCTTTTAAAAAGAGGGGGAGTTAAAGTATCTACAAAAAAT---CTAATCAAGCTATTTCAAATAATAGAACAATTTTGCCCATGGTTTCCAGAACAAGGAACTTTAGATCTAAAA---GATTGG---AAAAGA
A S Y L S F I K I L L K R G G V K V S T K N - L I K L F Q I I E Q F C P W F P E Q G T L D L K - D W - K R
K102
K103
K104
K106
K107
K108

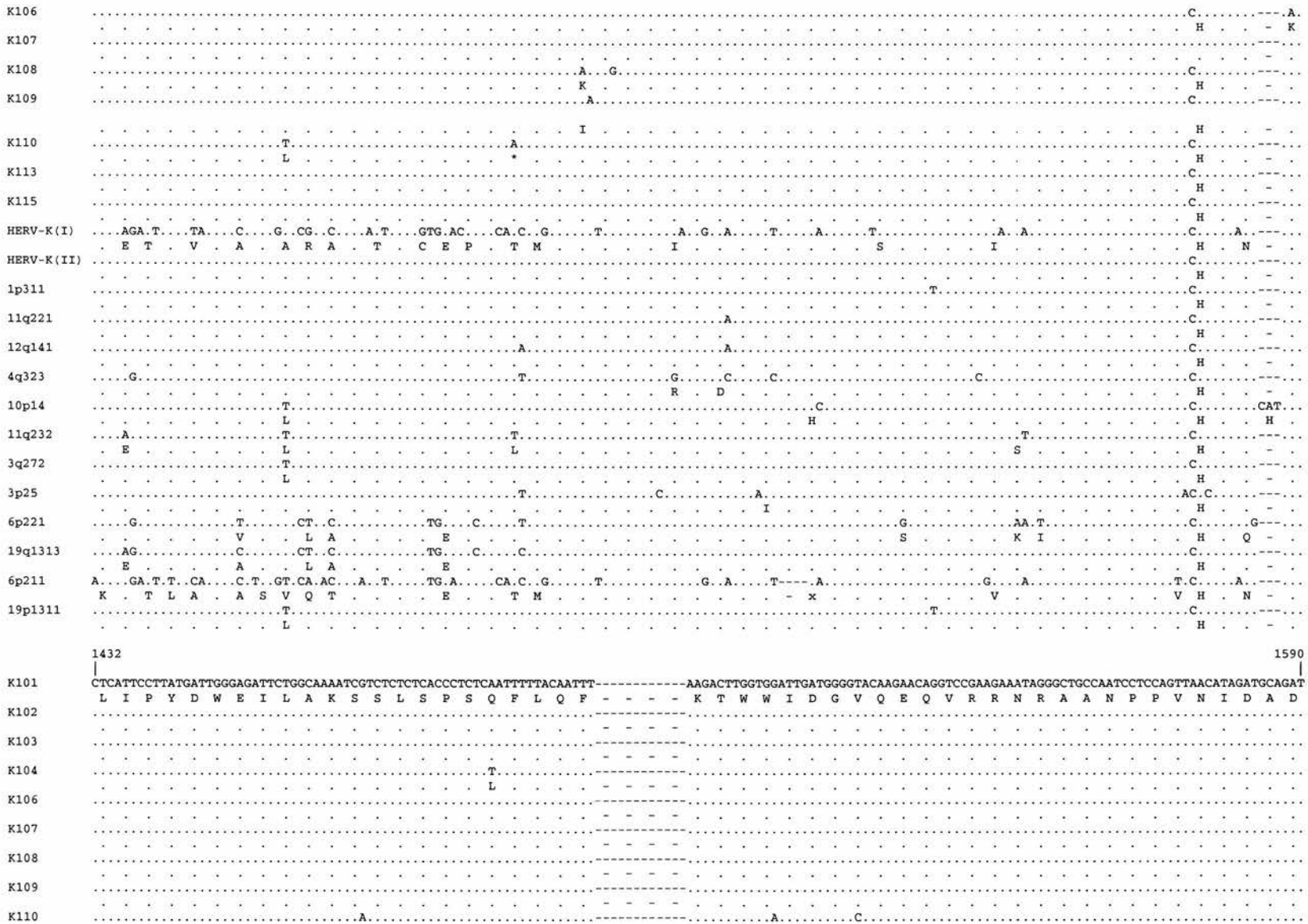

K109
K110
K113
K115
HERV-K(I)
HERV-K(II)
1p311
11q221
12q141
4q323
10p14
11q232
3q272
3p25
6p221
19q1313
6p211
19p1311
478
636
K101
ATTGGTAAGGAACTAAAACAAGCAGGTAGGAAGGGTAATATCATTCCACTTACAGTATGGAATGATTGGGCCATTATTAAAGCAGCTTTAGAACCATTTCAAACAGAAGAAGATAGCGTTTCAGTTTCTGATGCCCCTGGAAGCTGTTTAATAGATTGT
I G K E L K Q A G R K G N I I P L T V W N D W A I I K A A L E P F Q T E E D S V S V S D A P G S C L I D C
K102
K103
K104
K106
K107
K108
K109
K110
K113
K115
HERV-K(I)

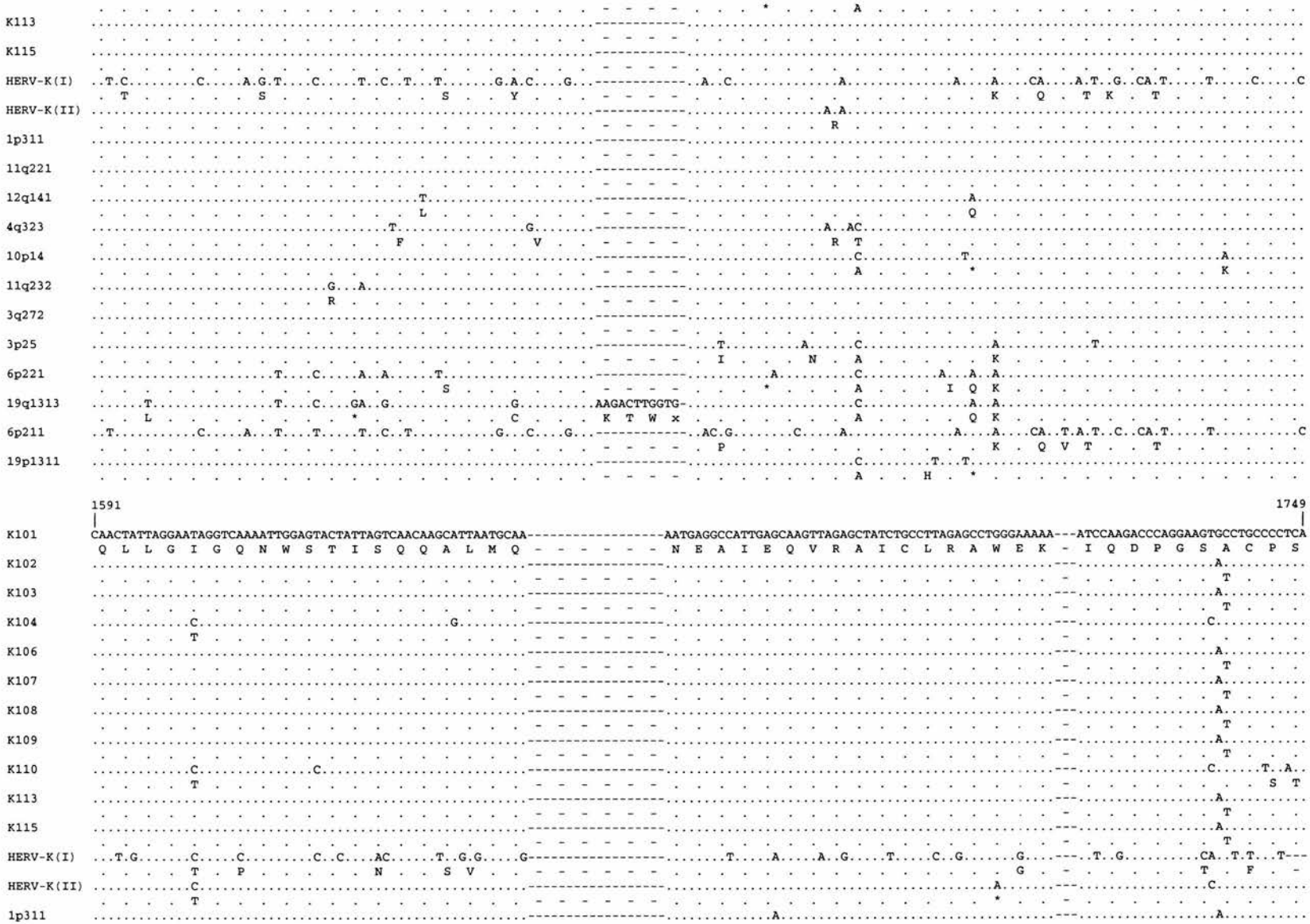

637
795
HERV-K(II)
1p311
11q221
12q141
4q323
10p14
11q232
3q272
3p25
6p221
19q1313
6p211
19p1311
K101
AATGAAAAGACAAGGAAAAAA---TCCCAGAAAGAAACGGAAAGTTTACATTGCGAATATGTAGCAGAGCCGGTAATGGCTCAGTCAACGCAAAATGTTGACTATAATCAA------TTACAGGAGGTGATATATCCTGAAACGTTAAAATTAGAAGGA
N E K T R K K - S Q K E T E S L H C E Y V A E P V M A Q S T Q N V D Y N Q - - L Q E V I Y P E T L K L E G
K102
K103
K104
K106
K107
K108
K109
K110
K113
K115
HERV-K(I)
HERV-K(II)
1p311
11q221
12q141
4q323

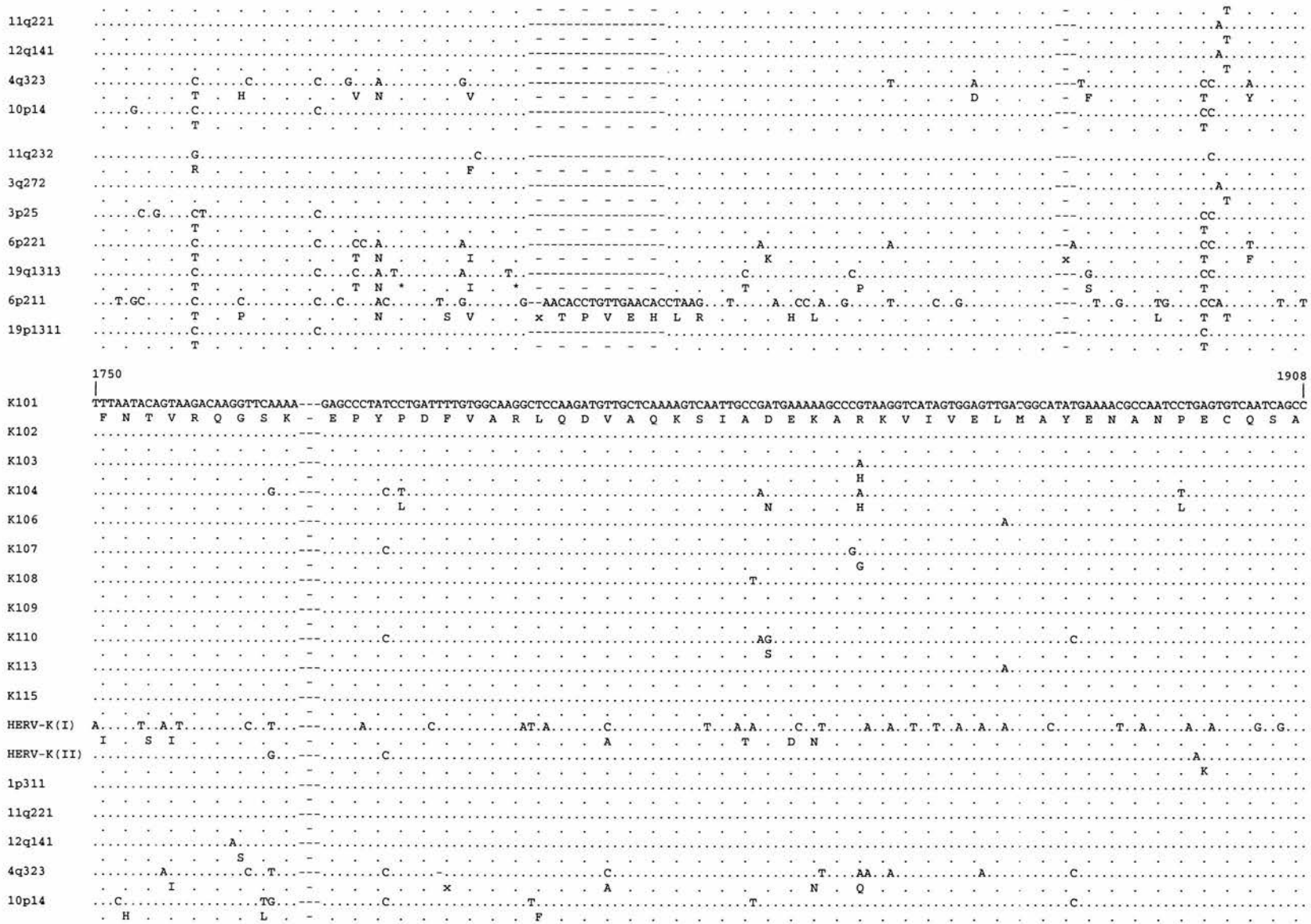

10p14
11q232
3q272
3p25
6p221
19q1313
6p211
19p1311
796
954
K101
K102
K103
K104
K106
K107
K108
K109
K110
K113
K115
HERV-K(I)
HERV-K(II)
1p311
11q221
12q141
4q323
10p14
11q232
3q272
3p25

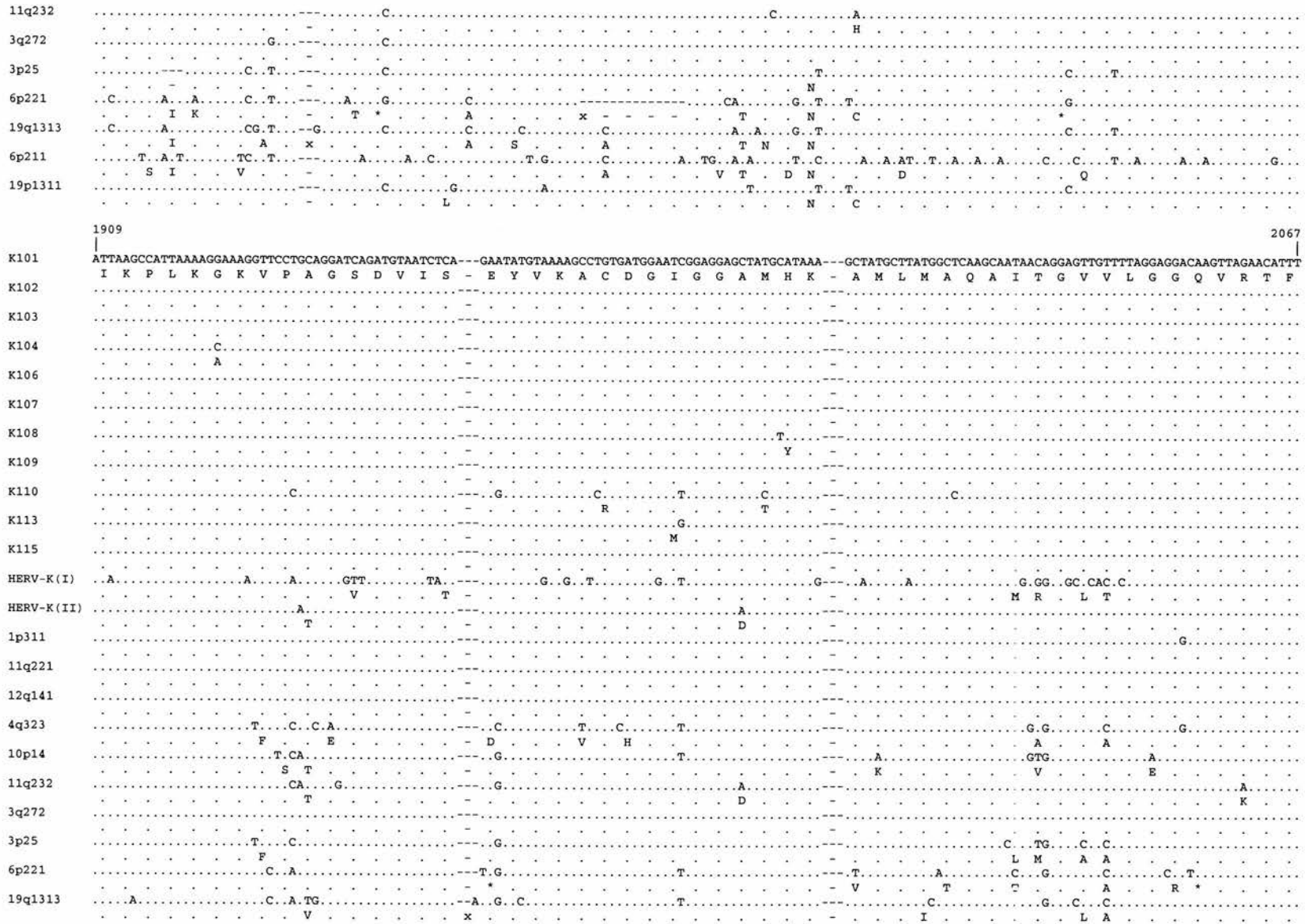

6p221
19q1313
6p211
19p1311
955
1113
K101
K102
K103
K104
K106
K107
K108
K109
K110
K113
K115
HERV-K(I)
HERV-K(II)
1p311
11q221
12q141
4q323
10p14
11q232
3q272
3p25
6p221
19q1313
6p211
19p1311

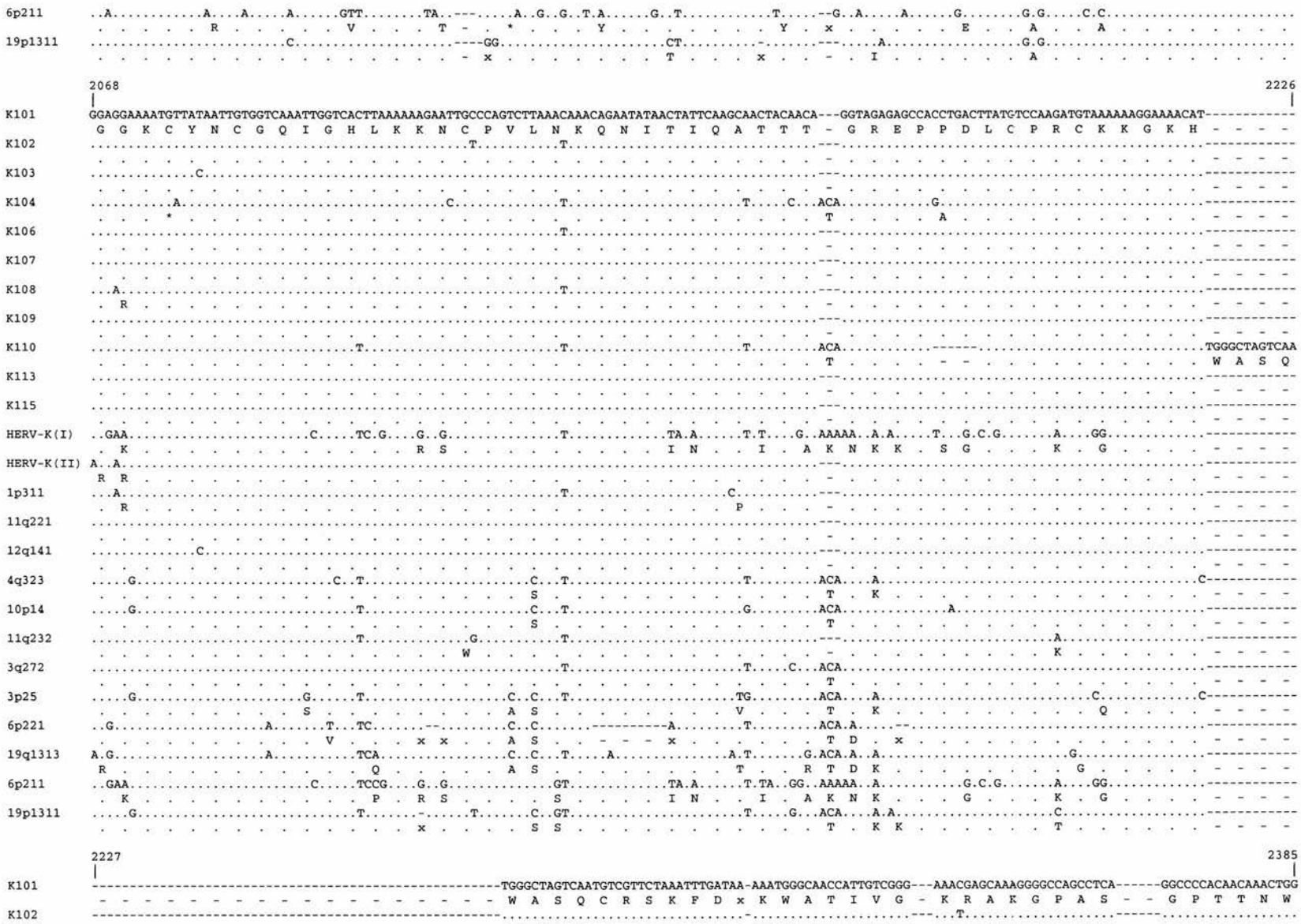

1114
1272
K101
AGGGCGCCATACCCTCAGCCGCCCACTAGGAGACTTAATCCTACGGCACCACCTAGTAGACAGGGTAGTGAATTACATGAAATTATTGATAAATCAAGAAAGGAAGGAGATACTGAGGCATGGCAATTCCCAGTAACGTTAGAACCGATGCCACCTGGA
R A P Y P Q P P T R R L N P T A P P S R Q G S E L H E I I D K S R K E G D T E A W Q F P V T L E P M P P G
K102
K103
K104
K106
K107
K108
K109
K110
K113
K115
HERV-K(I)
HERV-K(II)
1p311
11q221
12q141
4q323
10p14
11q232
3q272
3p25
6p221
19q1313
6p211
19p1311
1273
1431
K101
GAAGGAGCCCAAGAGGGAGAGCCTCCCACAGTTGAGGCCAGATACAAGTCTTTTTCGATAAAAATGCTAAAAGATATGAAAGAGGGAGTAAAACAGTATGGACCCAACTCCCCTTATATGAGGACATTATTAGATTCCATTGCTTATGGACAT---AGA
E G A Q E G E P P T V E A R Y K S F S I K M L K D M K E G V K Q Y G P N S P Y M R T L L D S I A Y G H - R
K102
K103
K104

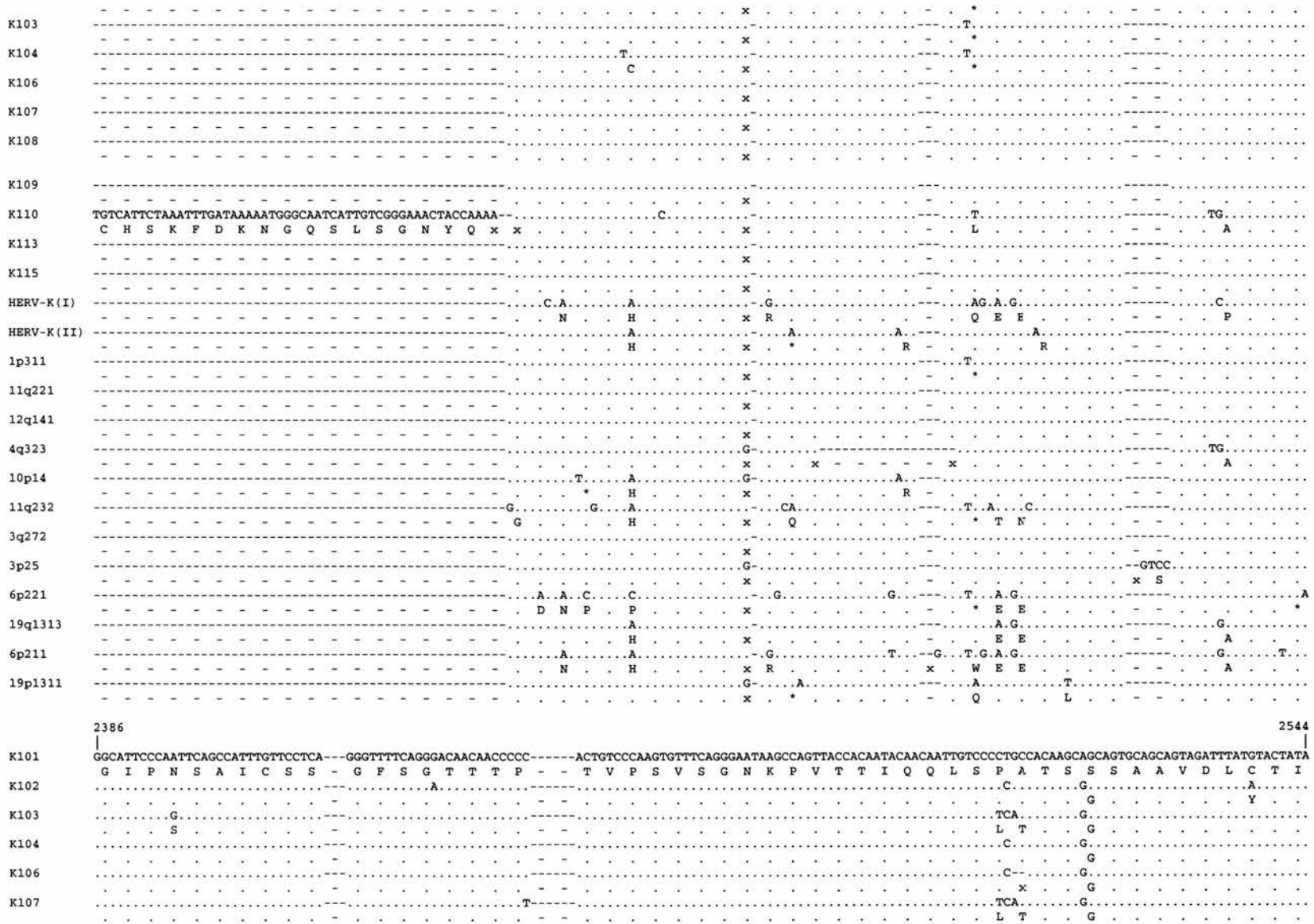

K106
K107
K108
K109
K110
K113
K115
HERV-K(I)
HERV-K(II)
1p311
11q221
12q141
4q323
10p14
11q232
3q272
3p25
6p221
19q1313
6p211
19p1311
1432
1590
CTCATTCCTTATGATTGGGAGATTCTGGCAAAATCGTCTCTCTCACCCTCTCAATTTTTACAATTT----------AAGACTTGGTGGATTGATGGGGTACAAGAACAGGTCCGAAGAAATAGGGCTGCCAATCCTCCAGTTAACATAGATGCAGAT
L I P Y D W E I L A K S S L S P S Q F L Q F - - - - K T W W I D G V Q E Q V R R N R A A N P P V N I D A D

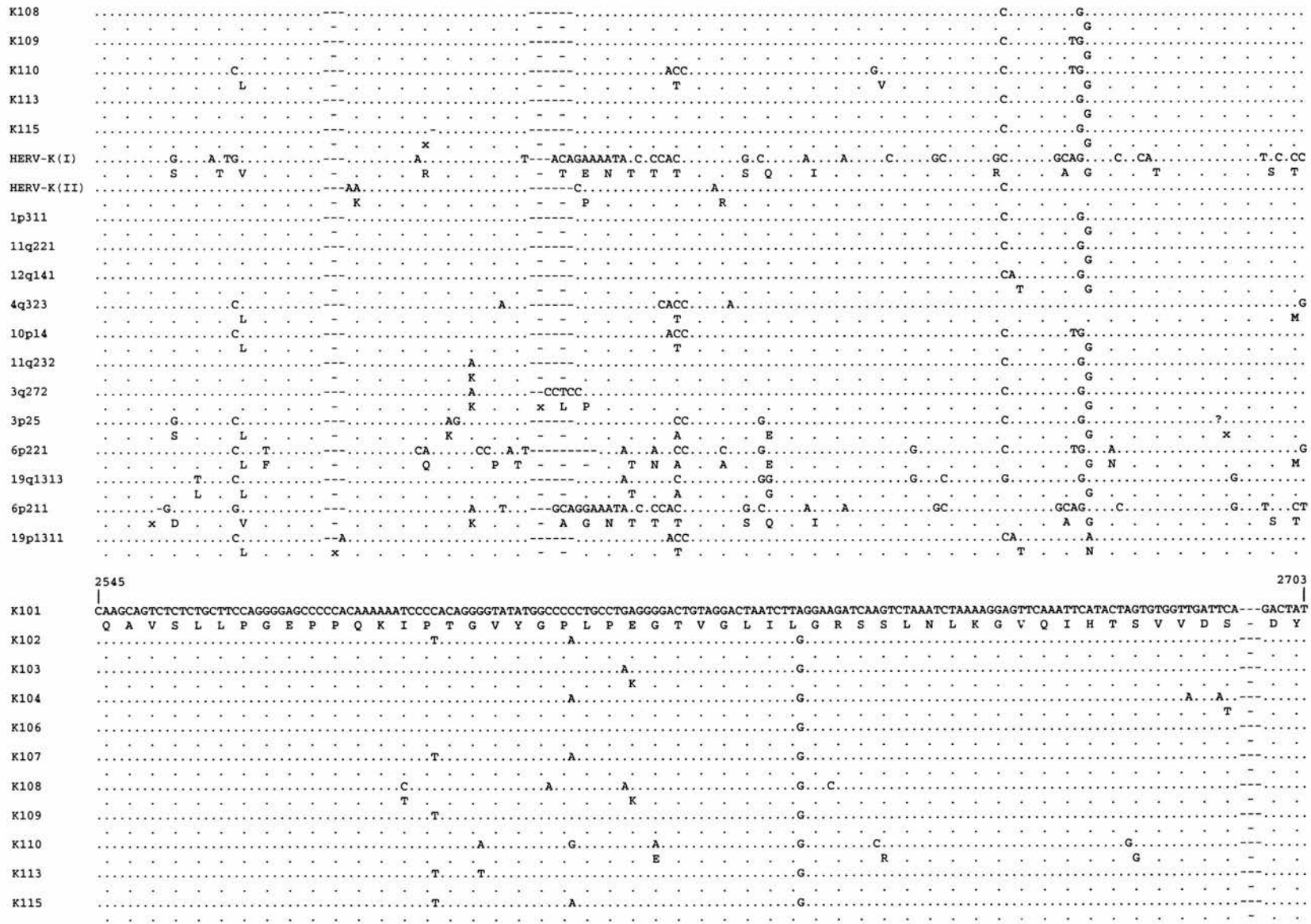

K113
K115
HERV-K(I)
HERV-K(II)
1p311
11q221
12q141
4q323
10p14
11q232
3q272
3p25
6p221
19q1313
6p211
19p1311
1591
1749
K101
CAACTATTAGGAATAGGTCAAAATTGGAGTACTATTAGTCAACAAGCATTAATGCAA----------AATGAGGCCATTGAGCAAGTTAGAGCTATCTGCCTTAGAGCCTGGGAAAAA---ATCCAAGACCCAGGAAGTGCCTGCCCCTCA
Q L L G I G Q N W S T I S Q Q A L M Q - - - - - - - N E A I E Q V R A I C L R A W E K - I Q D P G S A C P S
K102
K103
K104
K106
K107
K108
K109
K110
K113
K115
HERV-K(I)
HERV-K(II)
1p311

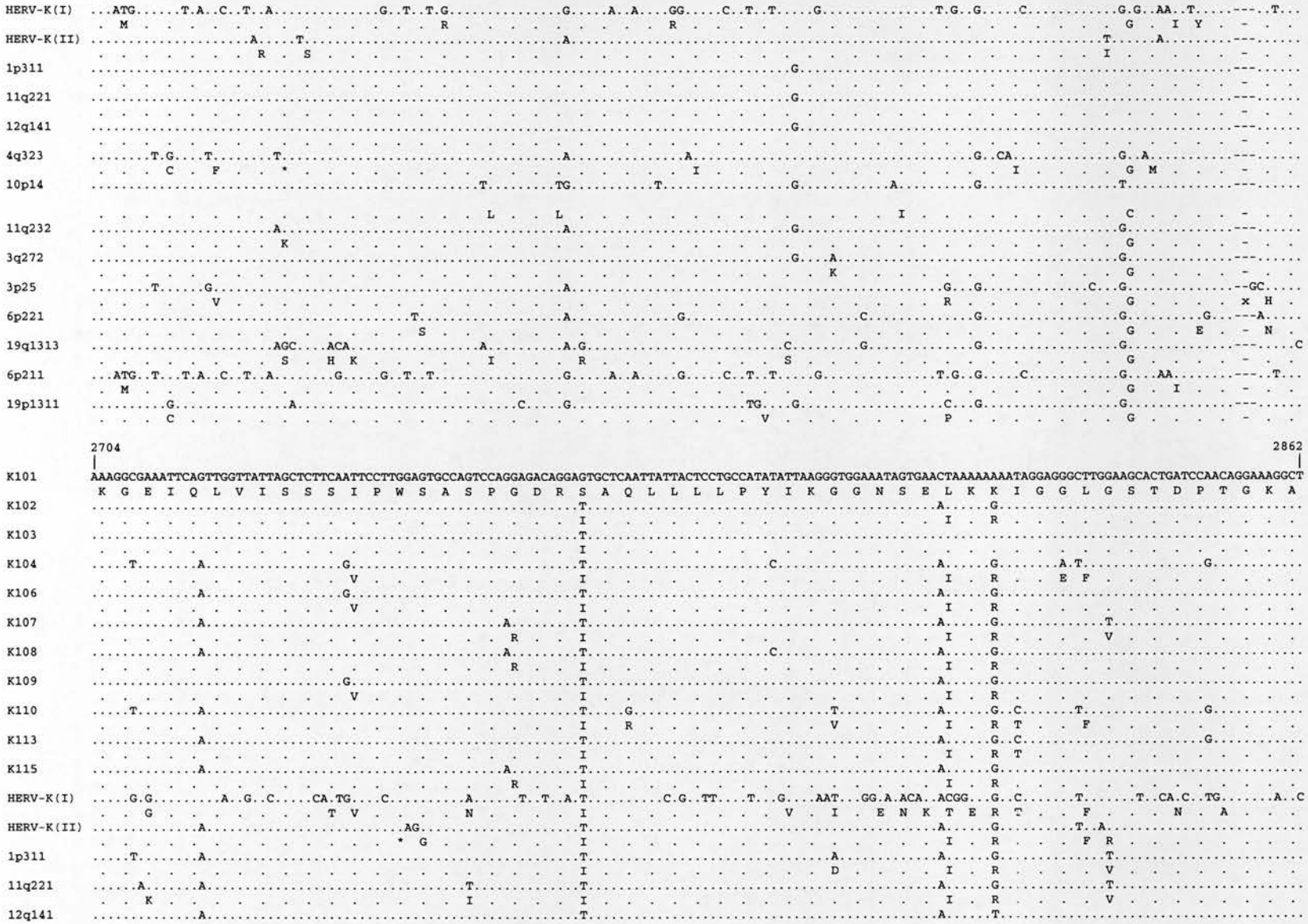

11q221
12q141
4q323
10p14
11q232
3q272
3p25
6p221
19q1313
6p211
19p1311
1750
1908
K101
TTTAATACAGTAAGACAAGGTTCAAAA---GAGCCCTATCCTGATTTTGTGGCAAGGCTCCAAGATGTTGCTCAAAAGTCAATTGCCGATGAAAAAGCCCGTAAGGTCATAGTGGAGTTGATGGCATATGAAAACGCCAATCCTGAGTGTCAATCAGCC
F N T V R Q G S K - E P Y P D F V A R L Q D V A Q K S I A D E K A R K V I V E L M A Y E N A N P E C Q S A
K102
K103
K104
K106
K107
K108
K109
K110
K113
K115
HERV-K(I)
HERV-K(II)
1p311
11q221
12q141
4q323
10p14

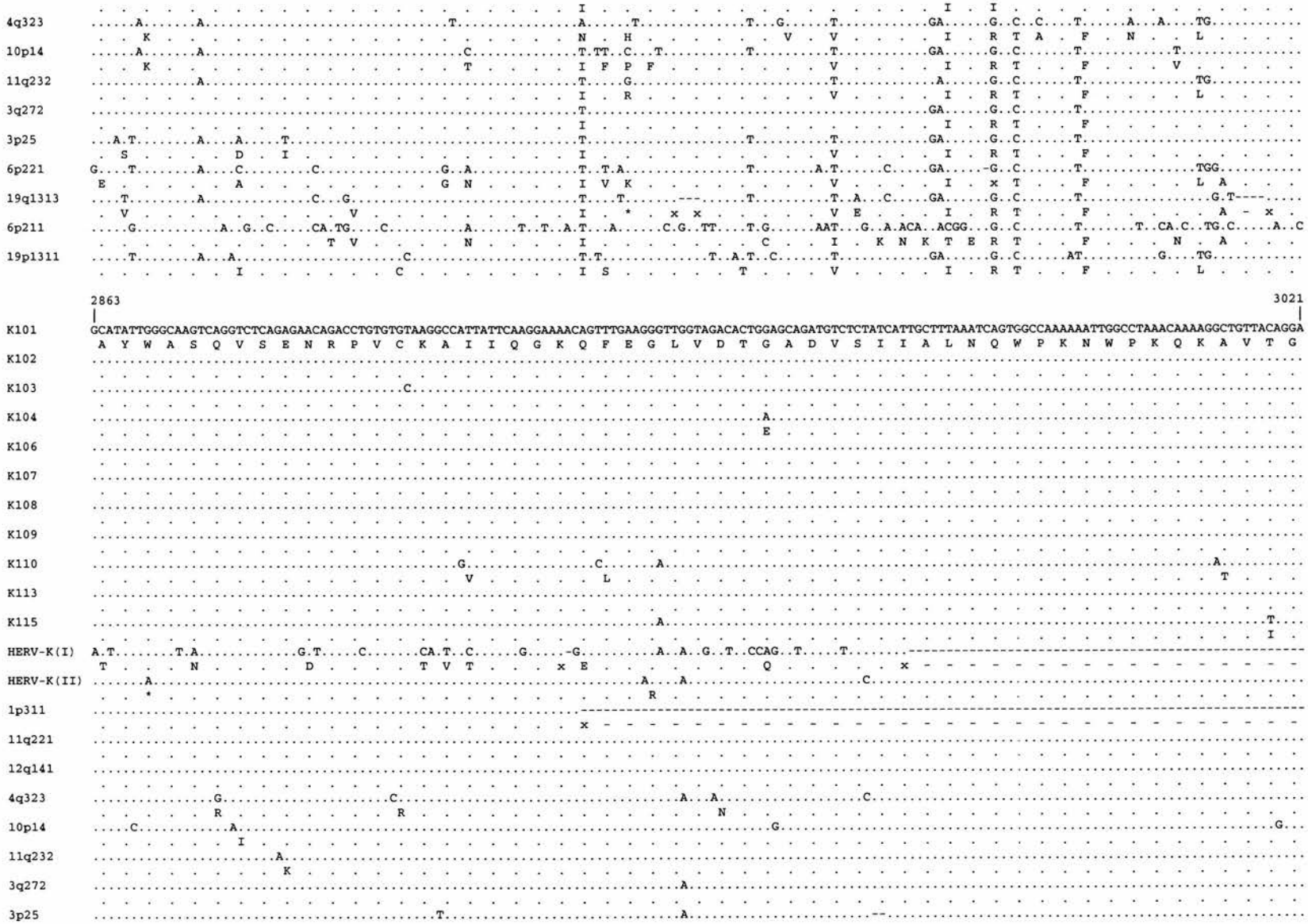

11q232
3q272
3p25
6p221
19q1313
6p211
19p1311
1909
2067
K101
ATTAAGCCATTAAAAGGAAAGGTTCCTGCAGGATCAGATGTAATCTCA---GAATATGTAAAAGCCTGTGATGGAATCGGAGGAGCTATGCATAAA---GCTATGCTTATGGCTCAAGCAATAACAGGAGTTGTTTTAGGAGGACAAGTTAGAACATTT
I K P L K G K V P A G S D V I S - E Y V K A C D G I G G A M H K - A M L M A Q A I T G V V L G G Q V R T F
K102
K103
K104
K106
K107
K108
K109
K110
K113
K115
HERV-K(I)
HERV-K(II)
1p311
11q221
12q141
4q323
10p14
11q232
3q272
3p25
6p221
19q1313

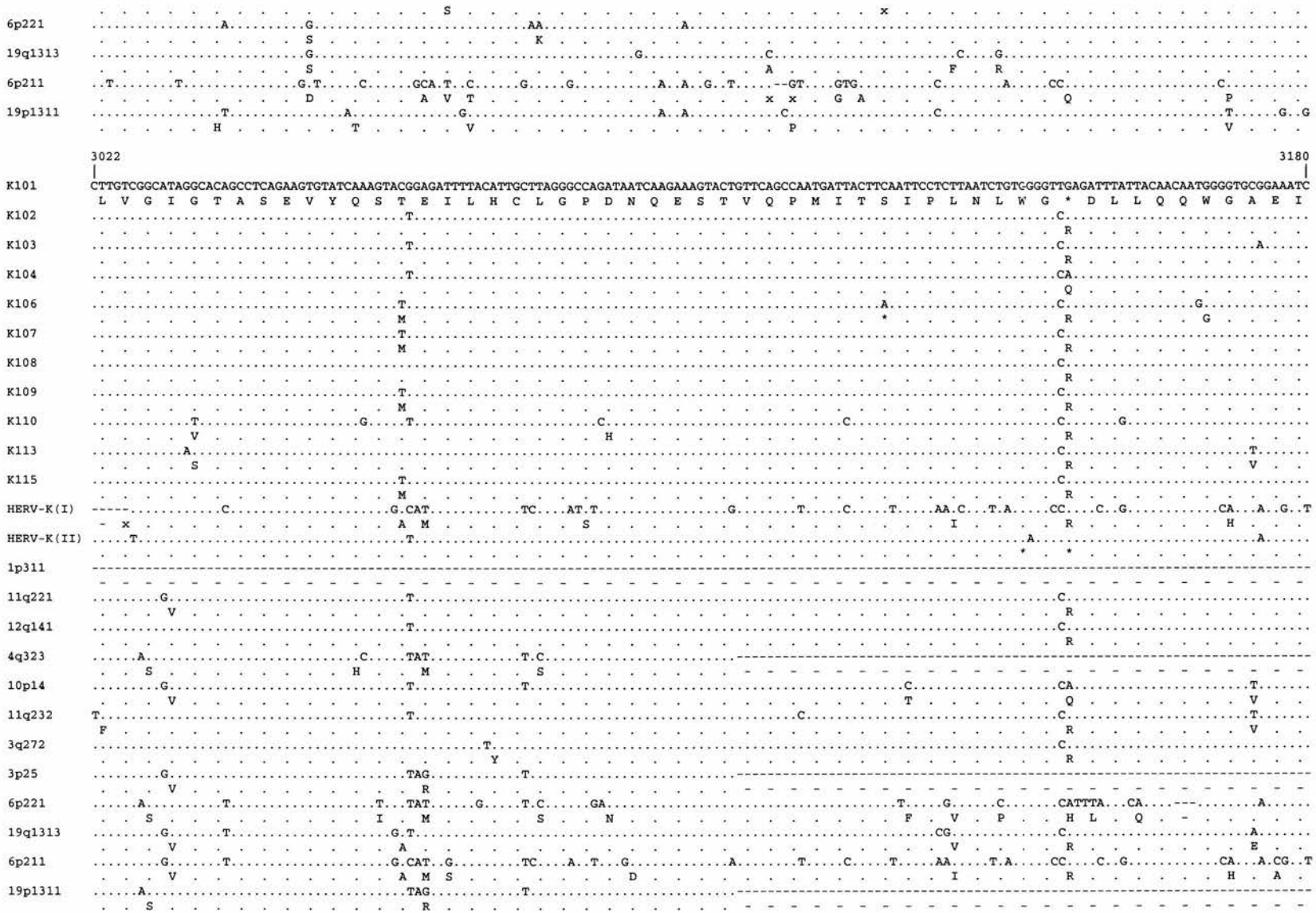

6p211
19p1311
2068
2226
K101
K102
K103
K104
K106
K107
K108
K109
K110
K113
K115
HERV-K(I)
HERV-K(II)
1p311
11q221
12q141
4q323
10p14
11q232
3q272
3p25
6p221
19q1313
6p211
19p1311
2227
2385
K101
K102

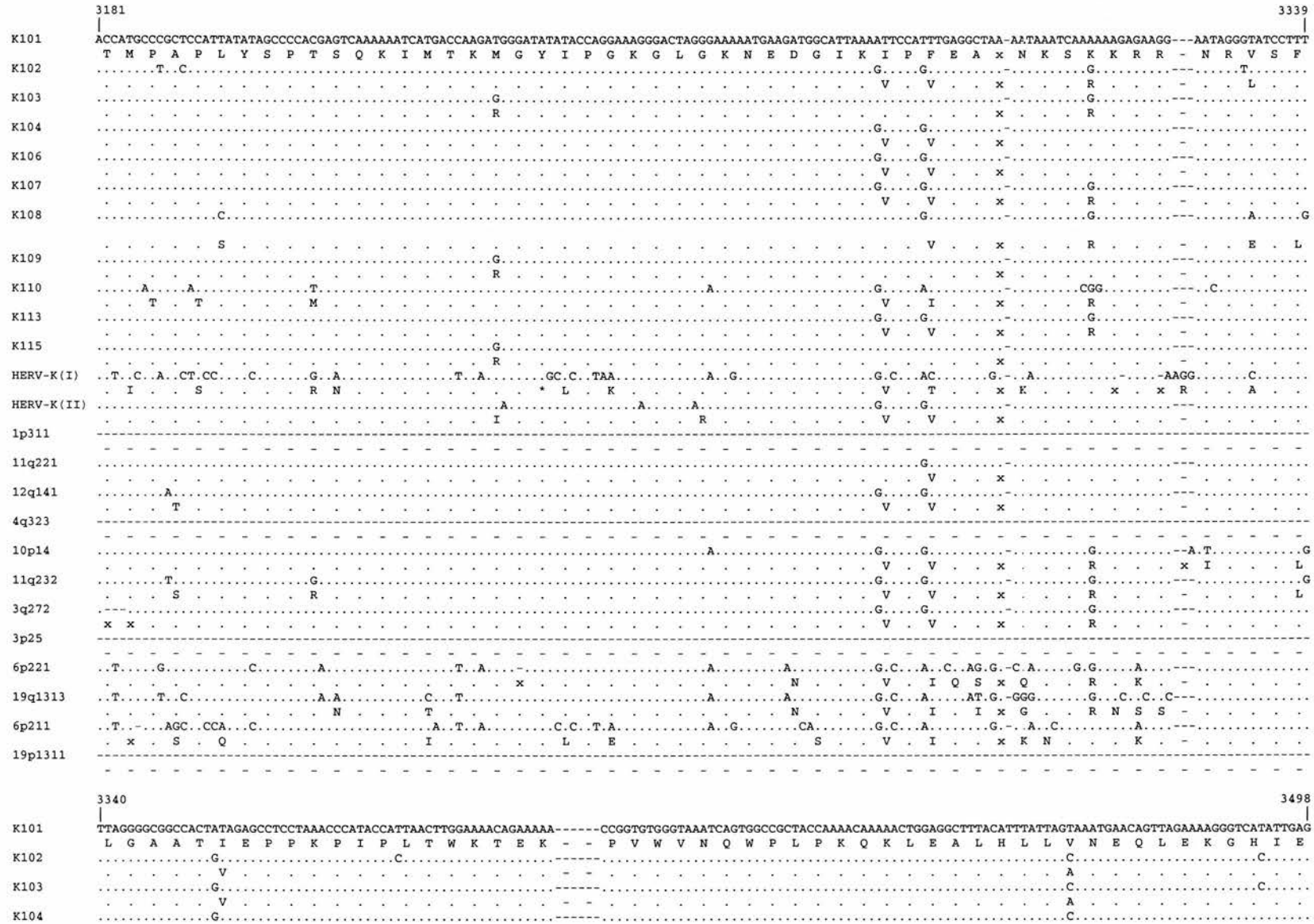

K103
K104
K106
K107
K108
K109
K110
K113
K115
HERV-K(I)
HERV-K(II)
1p311
11q221
12q141
4q323
10p14
11q232
3q272
3p25
6p221
19q1313
6p211
19p1311
2386
2544
K101
K102
K103
K104
K106
K107

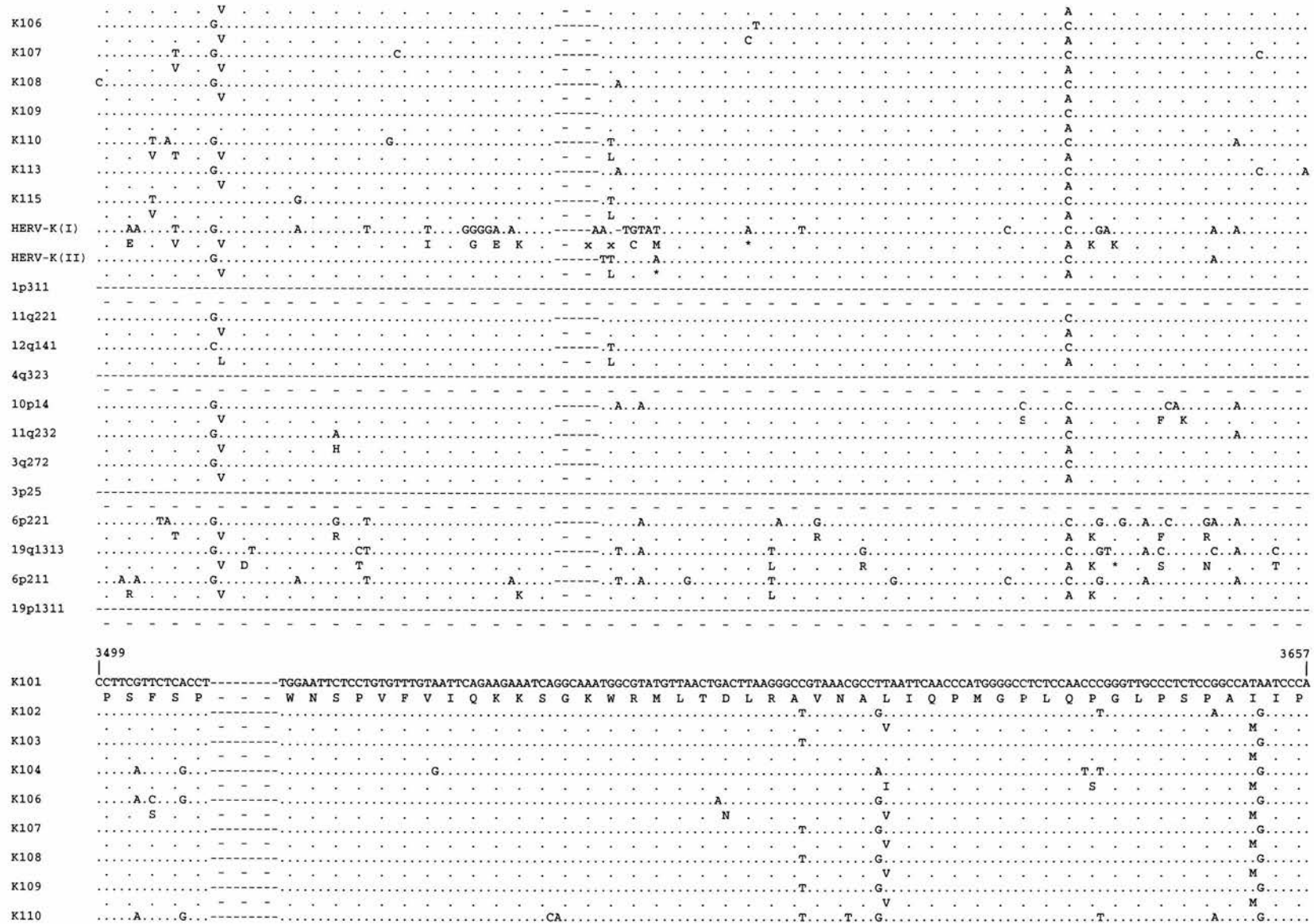

K108
K109
K110
K113
K115
HERV-K(I)
HERV-K(II)
1p311
11q221
12q141
4q323
10p14
11q232
3q272
3p25
6p221
19q1313
6p211
19p1311
2545
2703
K101
CAAGCAGTCTCTCTGCTTCCAGGGGAGCCCCCACAAAAAATCCCCACAGGGGTATATGGCCCCCTGCCTGAGGGGACTGTAGGACTAATCTTAGGAAGATCAAGTCTAAATCTAAAAGGAGTTCAAATTCATACTAGTGTGGTTGATTCA---GACTAT
Q A V S L L P G E P P Q K I P T G V Y G P L P E G T V G L I L G R S S L N L K G V Q I H T S V V D S - D Y
K102
K103
K104
K106
K107
K108
K109
K110
K113
K115

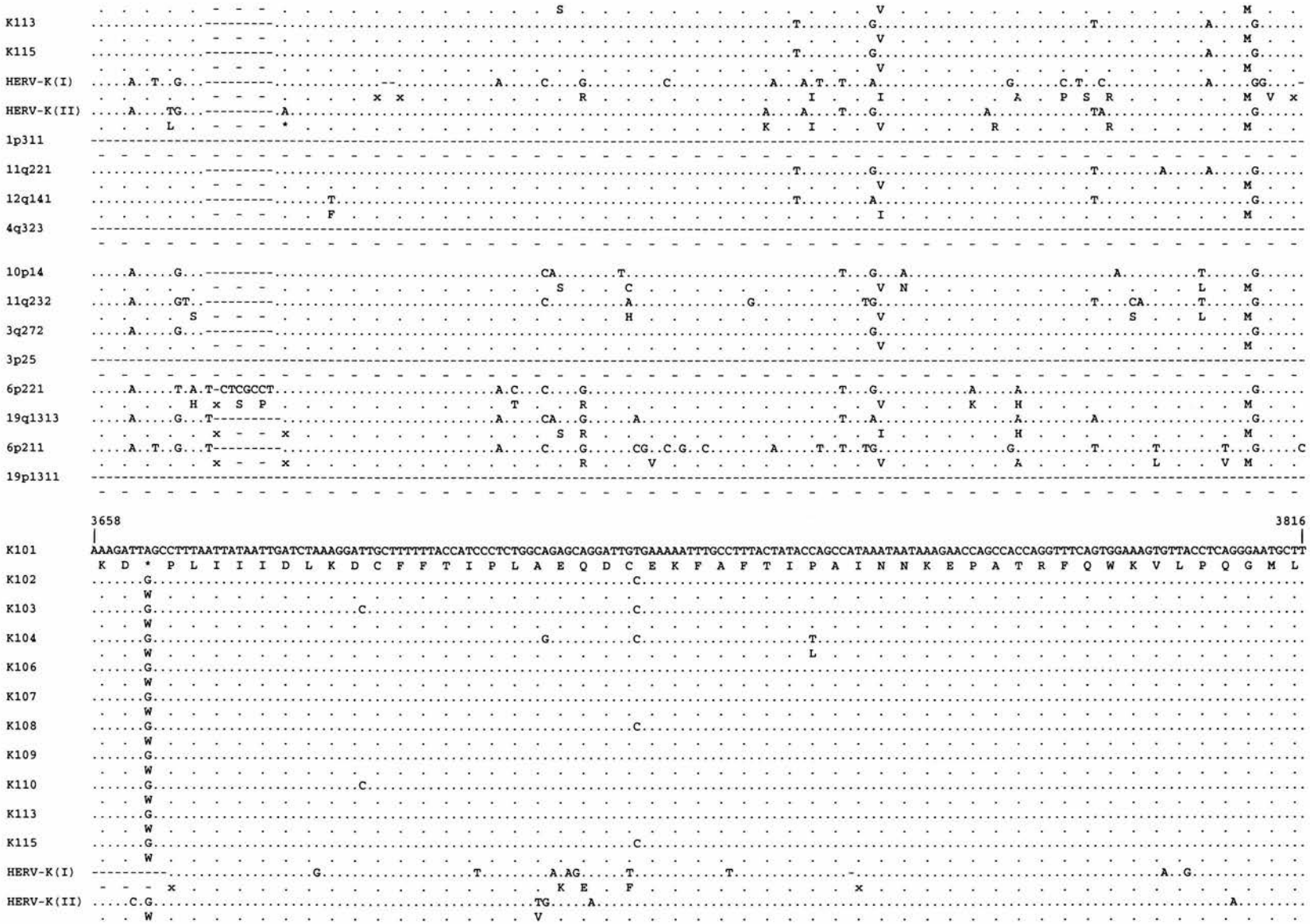

HERV-K(I)
HERV-K(II)
1p311
11q221
12q141
4q323
10p14
11q232
3q272
3p25
6p221
19q1313
6p211
19p1311
2704
2862
K101
K102
K103
K104
K106
K107
K108
K109
K110
K113
K115
HERV-K(I)
HERV-K(II)
1p311
11q221
12q141

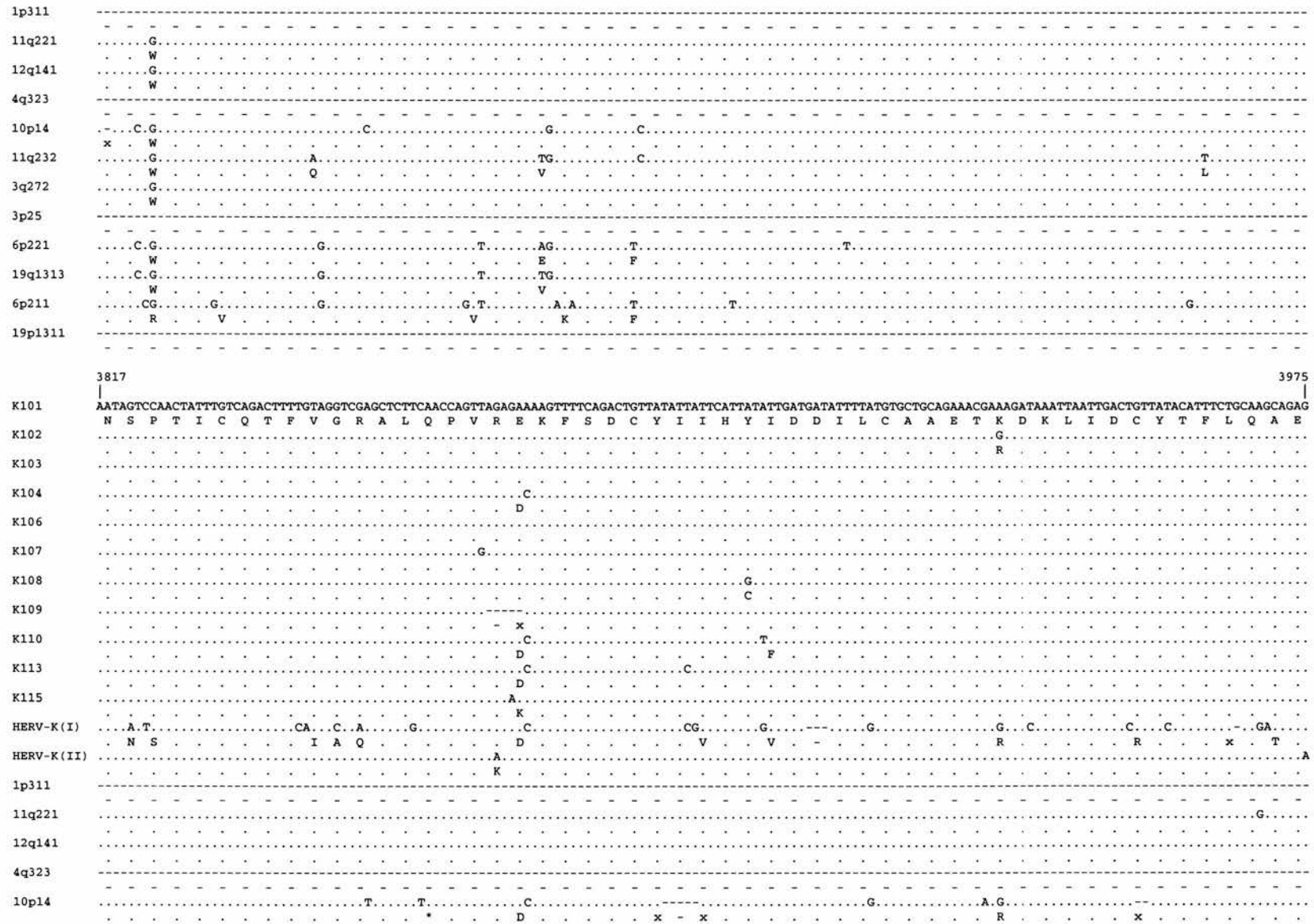

4q323
10p14
11q232
3q272
3p25
6p221
19q1313
6p211
19p1311
2863
3021
K101
GCATATTGGGCAAGTCAGGTCTCAGAGAACAGACCTGTGTGTAAGGCCATTATTCAAGGAAAACAGTTTGAAGGGTTGGTAGACACTGGAGCAGATGTCTCTATCATTGCTTTAAATCAGTGGCCAAAAAATTGGCCTAAACAAAAGGCTGTTACAGGA
A Y W A S Q V S E N R P V C K A I I Q G K Q F E G L V D T G A D V S I I A L N Q W P K N W P K Q K A V T G
K102
K103
K104
K106
K107
K108
K109
K110
K113
K115
HERV-K(I)
HERV-K(II)
1p311
11q221
12q141
4q323
10p14
11q232
3q272
3p25

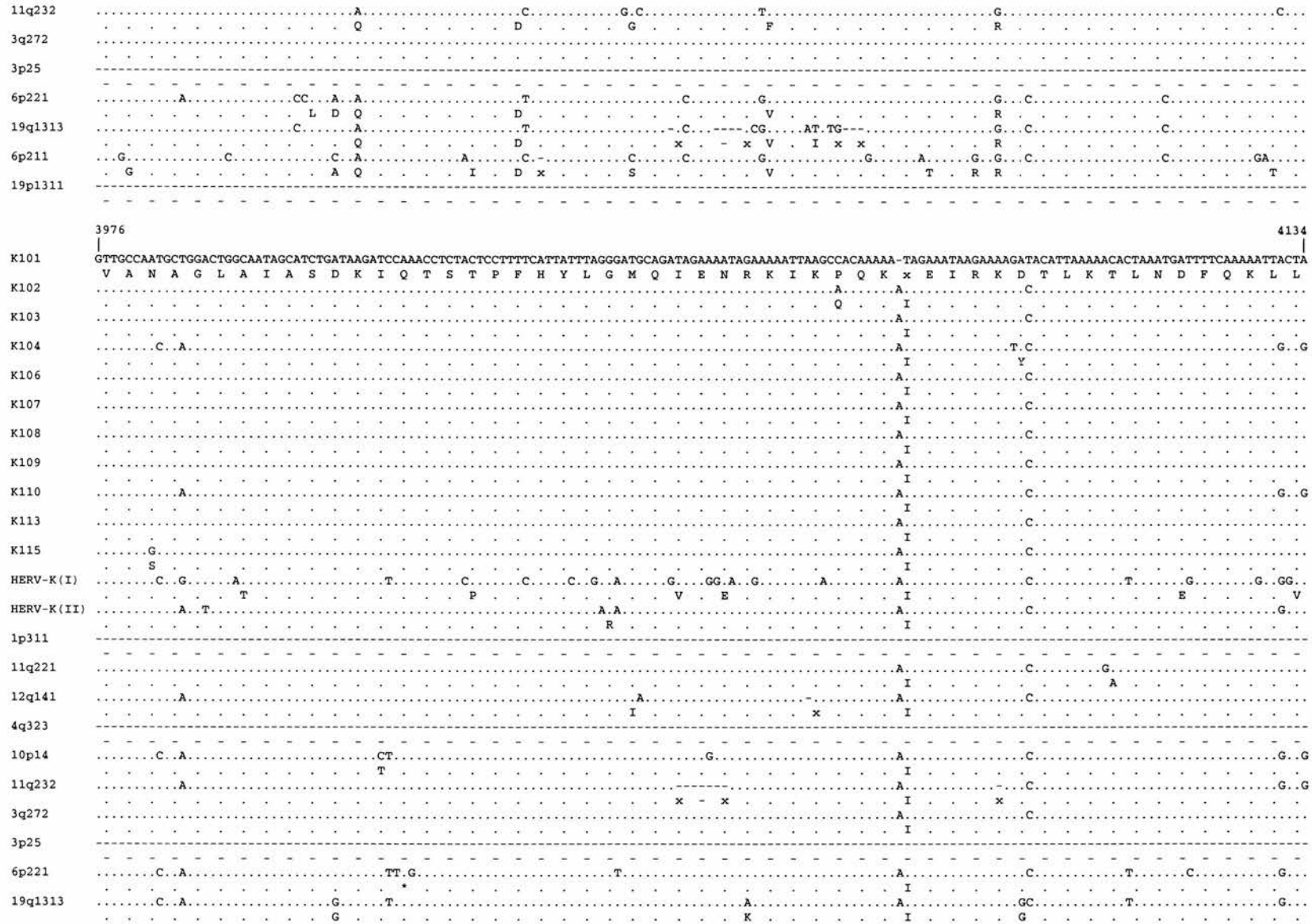

3022
3180
6p221
19q1313
6p211
19p1311
K101
K102
K103
K104
K106
K107
K108
K109
K110
K113
K115
HERV-K(I)
HERV-K(II)
1p311
11q221
12q141
4q323
10p14
11q232
3q272
3p25
6p221
19q1313
6p211
19p1311

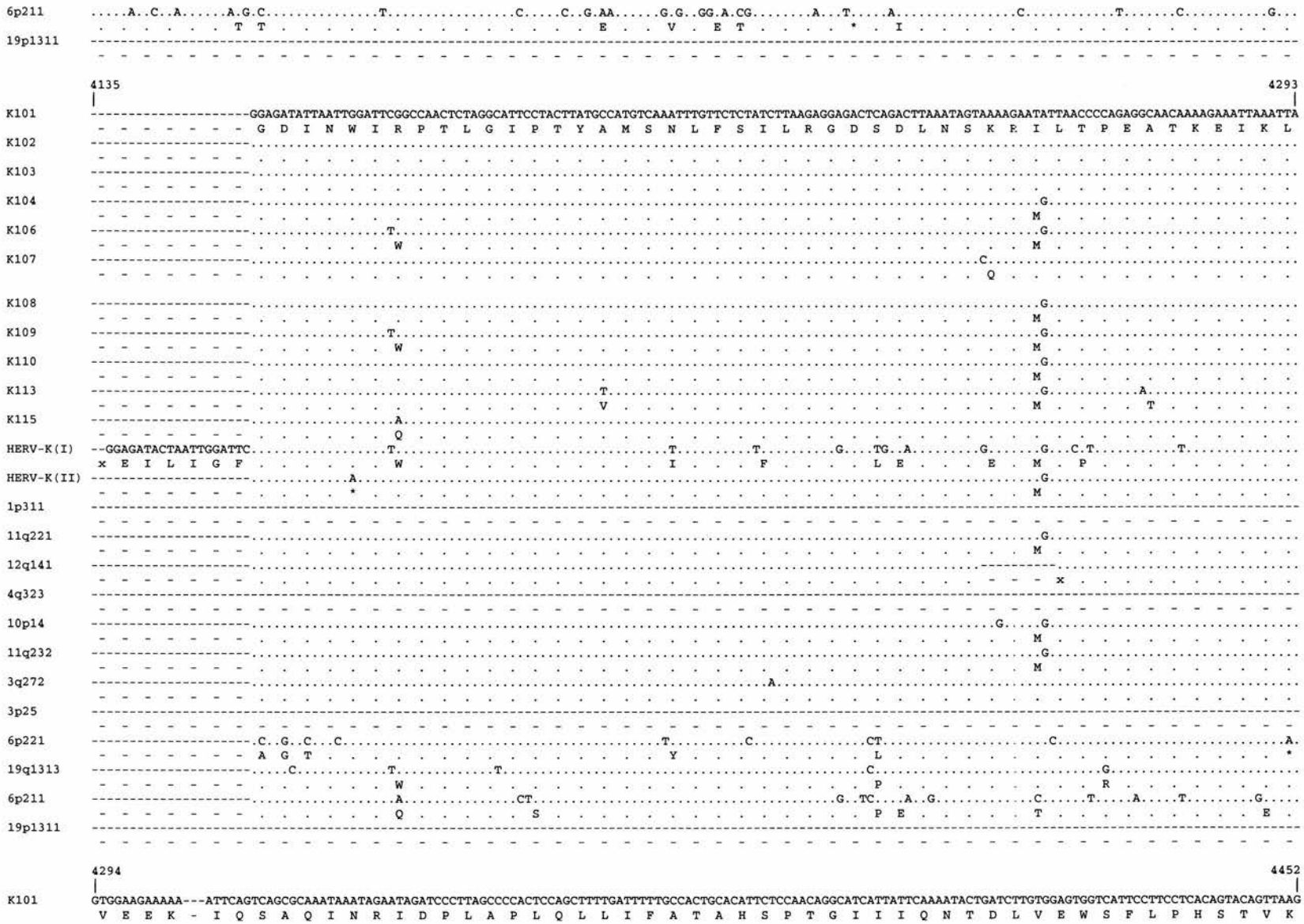

3181
3339
K101
ACCATGCCCGCTCCATTATATAGCCCCACGAGTCAAAAAATCATGACCAAGATGGGATATATACCAGGAAAGGGACTAGGGAAAAATGAAGATGGCATTAAAATTCCATTTGAGGCTAA-AATAAATCAAAAAAGAGAAGG---AATAGGGTATCCTTT
T M P A P L Y S P T S Q K I M T K M G Y I P G K G L G K N E D G I K I P F E A x N K S K K R R - N R V S F
K102
K103
K104
K106
K107
K108
K109
K110
K113
K115
HERV-K(I)
HERV-K(II)
1p311
11q221
12q141
4q323
10p14
11q232
3q272
3p25
6p221
19q1313
6p211
19p1311
3340
3498
K101
TTAGGGGCGGCCACTATAGAGCCTCCTAAACCCATACCATTAACTTGGAAAACAGAAAAA------CCGGTGTGGGTAAATCAGTGGCCGCTACCAAAACAAAAACTGGAGGCTTTACATTTATTAGTAAATGAACAGTTAGAAAAGGGTCATATTGAG
L G A A T I E P P K P I P L T W K T E K - - P V W V N Q W P L P K Q K L E A L H L L V N E Q L E K G H I E
K102
K103
K104

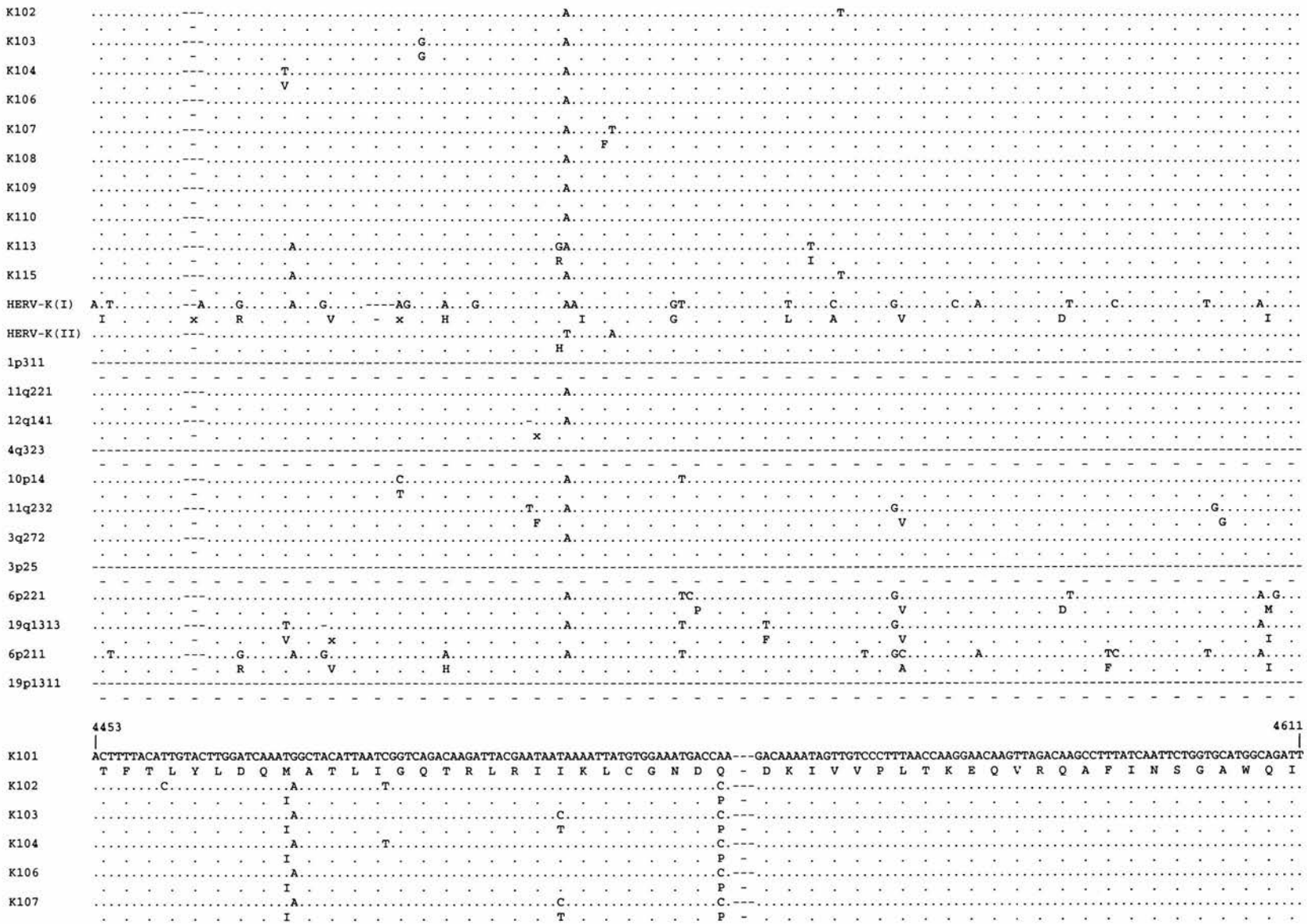

K106
K107
K108
K109
K110
K113
K115
HERV-K(I)
HERV-K(II)
1p311
11q221
12q141
4q323
10p14
11q232
3q272
3p25
6p221
19q1313
6p211
19p1311
3499
3657
K101
K102
K103
K104
K106
K107
K108
K109
K110

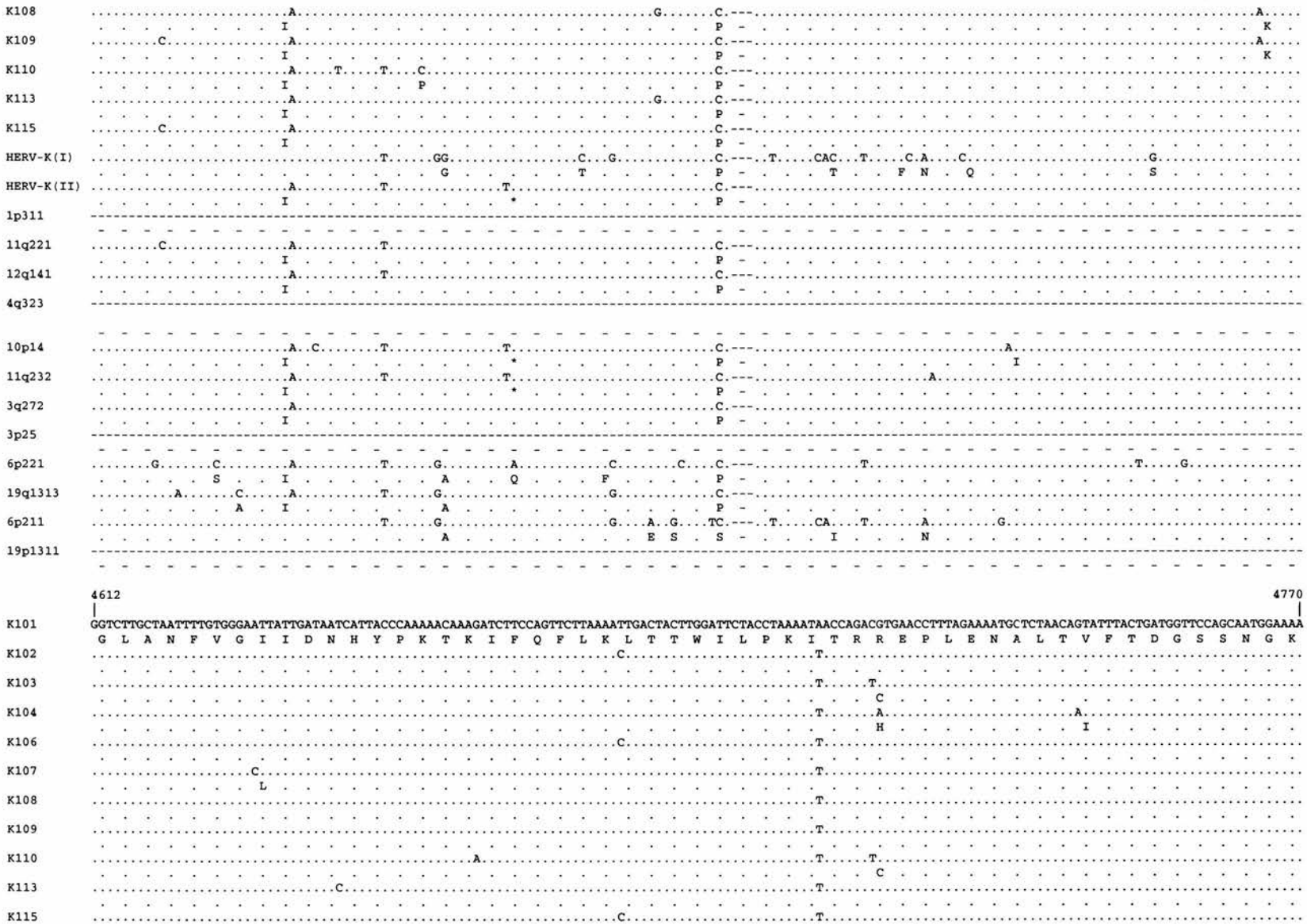

3658
3816
AAAGATTAGCCTTTAATTATAATTGATCTAAAGGATTGCTTTTTTACCATCCCTCTGGCAGAGCAGGATTGTGAAAAATTTGCCTTTACTATACCAGCCATAAATAATAAAGAACCAGCCACCAGGTTTCAGTGGAAAGTGTTACCTCAGGGAATGCTT
K D * P L I I I D L K D C F F T I P L A E Q D C E K F A F T I P A I N N K E P A T R F Q W K V L P Q G M L
K101
K102
K103
K104
K106
K107
K108
K109
K110
K113
K115
HERV-K(I)
HERV-K(II)
1p311
1q221
12q141
4q323
10p14
11q232
3q272
3p25
6p221
19q1313
6p211
19p1311

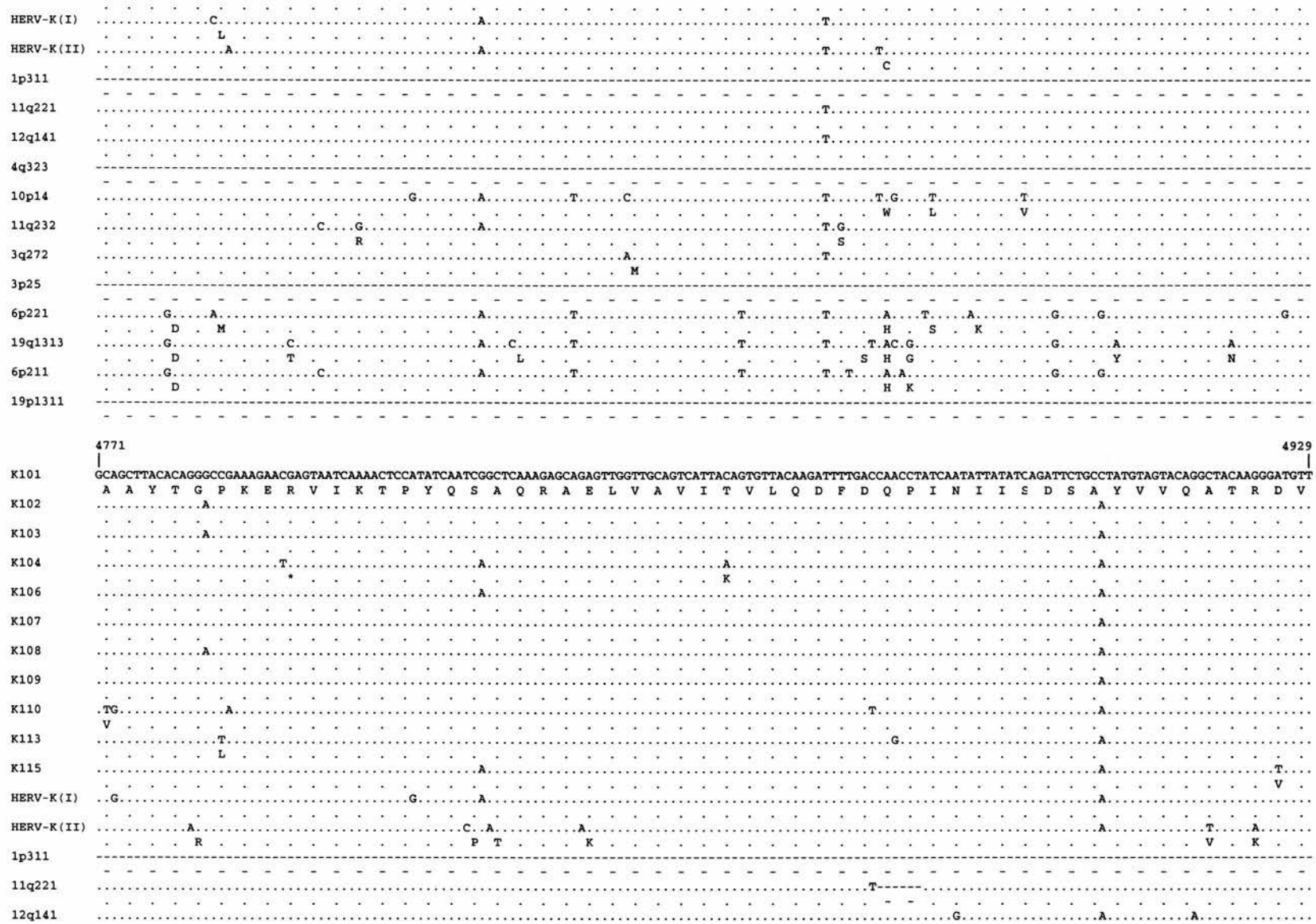

1p311
11q221
12q141
4q323
10p14
11q232
3q272
3p25
6p221
19q1313
6p211
19p1311
3817
3975
K101
AATAGTCCAACTATTTGTCAGACTTTTGTAGGTCGAGCTCTTCAACCAGTTAGAGAAAAGTTTTCAGACTGTTATATTATTCATTATATTGATGATATTTTATGTGCTGCAGAAACGAAAGATAAATTAATTGACTGTTATACATTTCTGCAAGCAGAG
N S P T I C Q T F V G R A L Q P V R E K F S D C Y I I H Y I D D I L C A A E T K D K L I D C Y T F L Q A E
K102
K103
K104
K106
K107
K108
K109
K110
K113
K115
HERV-K(I)
HERV-K(II)
1p311
11q221
12q141
4q323
10p14

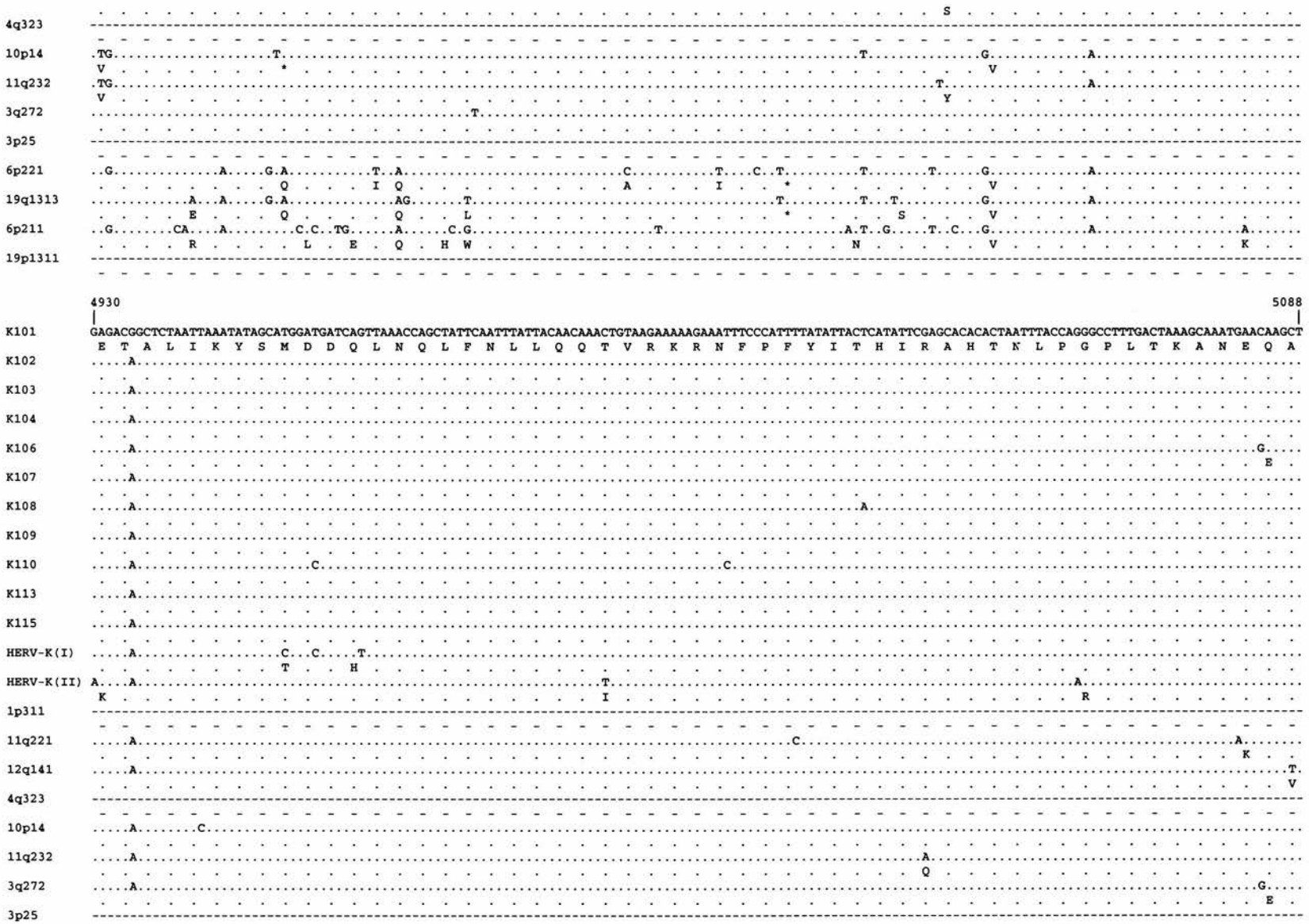

11q232
3q272
3p25
6p221
19q1313
6p211
19p1311
3976
4134
K101
GTTGCCAATGCTGGACTGGCAATAGCATCTGATAAGATCCAAACCTCTACTCCTTTTCATTATTTAGGGATGCAGATAGAAAATAGAAAAATTAAGCCACAAAAA-TAGAAATAAGAAAAGATACATTAAAAACACTAAATGATTTTCAAAAATTACTA
V A N A G L A I A S D K I Q T S T P F H Y L G M Q I E N R K I K P Q K x E I R K D T L K T L N D F Q K L L
K102
K103
K104
K106
K107
K108
K109
K110
K113
K115
HERV-K(I)
HERV-K(II)
1p311
11q221
12q141
4q323
10p14
11q232
3q272
3p25
6p221
19q1313

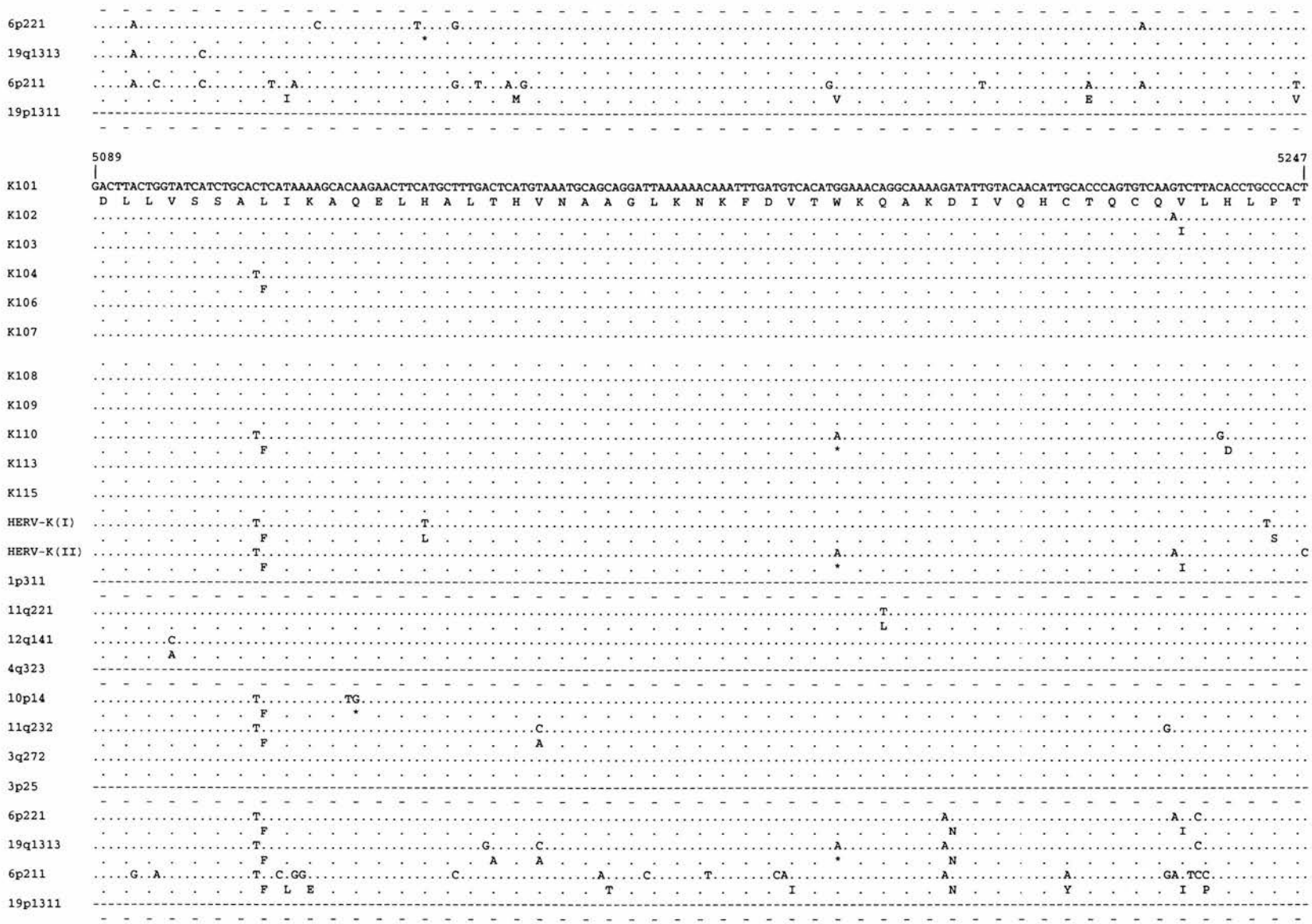

6p211
19p1311
4135
4293
K101
K102
K103
K104
K106
K107
K108
K109
K110
K113
K115
HERV-K(I)
HERV-K(II)
1p311
11q221
12q141
4q323
10p14
11q232
3q272
3p25
6p221
19q1313
6p211
19p1311
4294
4452
K101

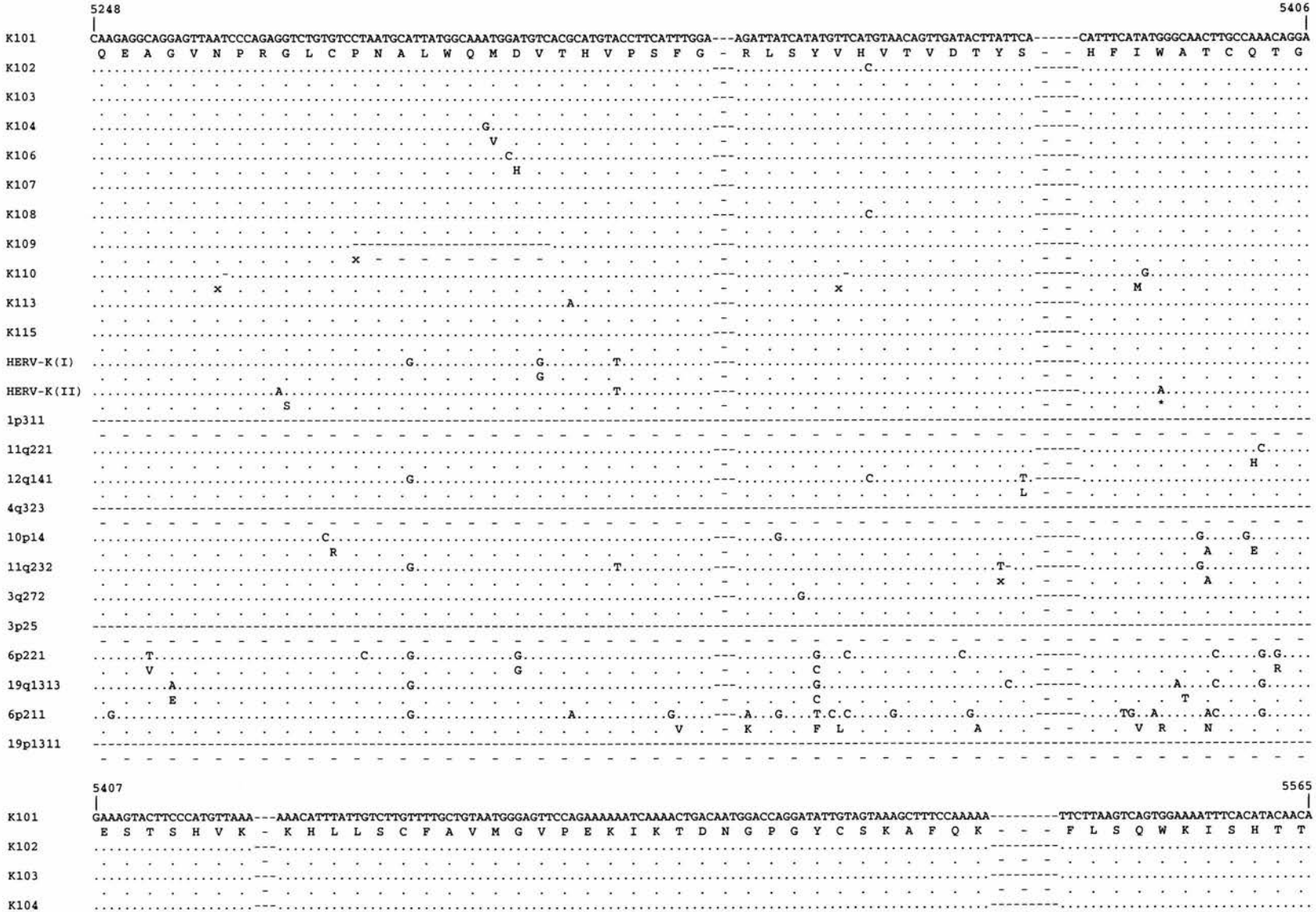

K102
K103
K104
K106
K107
K108
K109
K110
K113
K115
HERV-K(I)
HERV-K(II)
1p311
11q221
12q141
4q323
10p14
11q232
3q272
3p25
6p221
19q1313
6p211
19p1311
4453
4611
K101
ACTTTTACATTGTACTTGGATCAAATGGCTACATTAATCGGTCAGACAAGATTACGAATAATAAAATTATGTGGAAATGACCAA---GACAAAATAGTTGTCCCTTTAACCAAGGAACAAGTTAGACAAGCCTTTATCAATTCTGGTGCATGGCAGATT
T F T L Y L D Q M A T L I G Q T R L R I I K L C G N D Q - D K I V V P L T K E Q V R Q A F I N S G A W Q I
K102
K103
K104
K106
K107

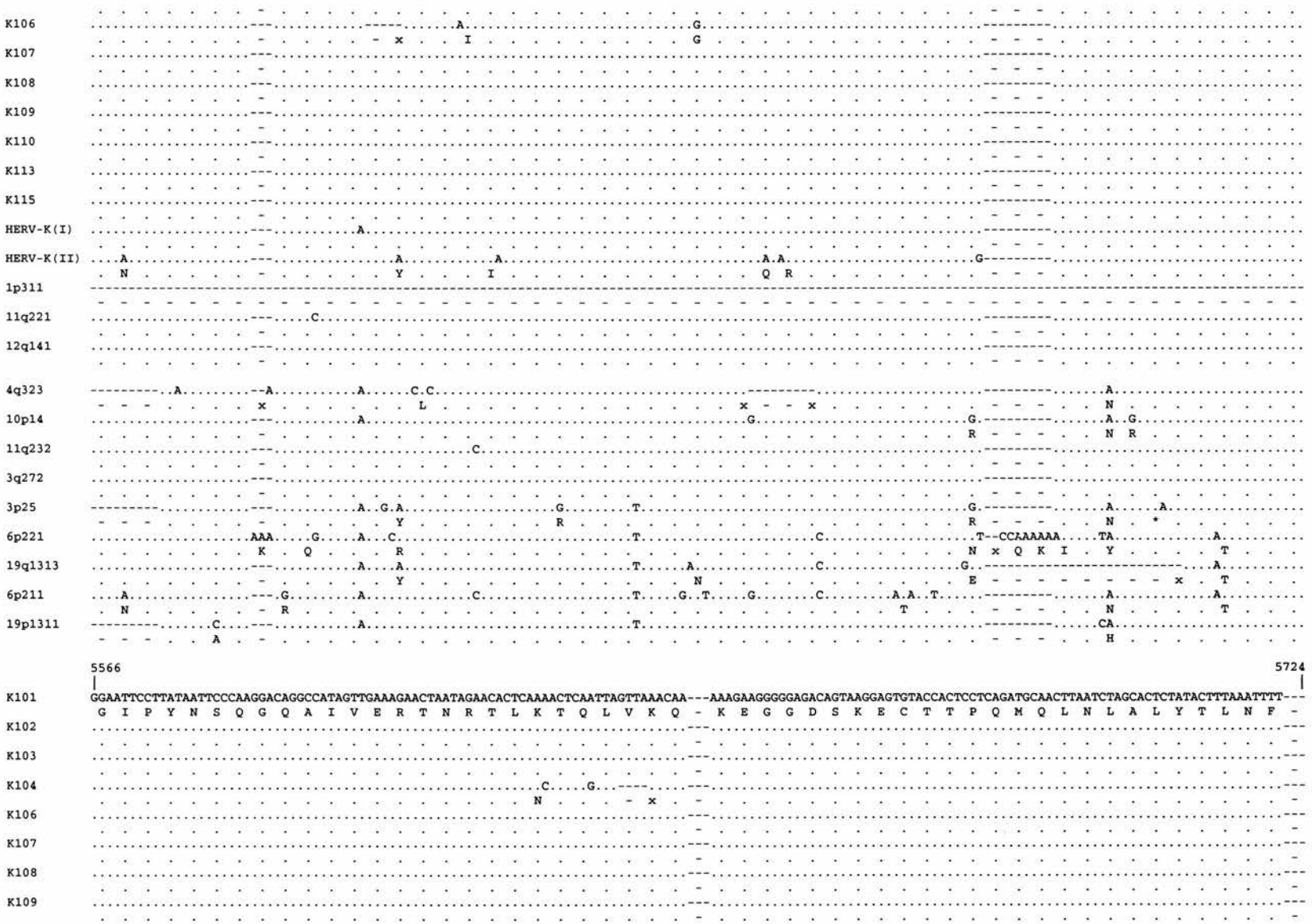

K108
K109
K110
K113
K115
HERV-K(I)
HERV-K(II)
1p311
11q221
12q141
4q323
10p14
11q232
3q272
3p25
6p221
19q1313
6p211
19p1311
4612
4770
K101
K102
K103
K104
K106
K107
K108
K109
K110
K113
K115

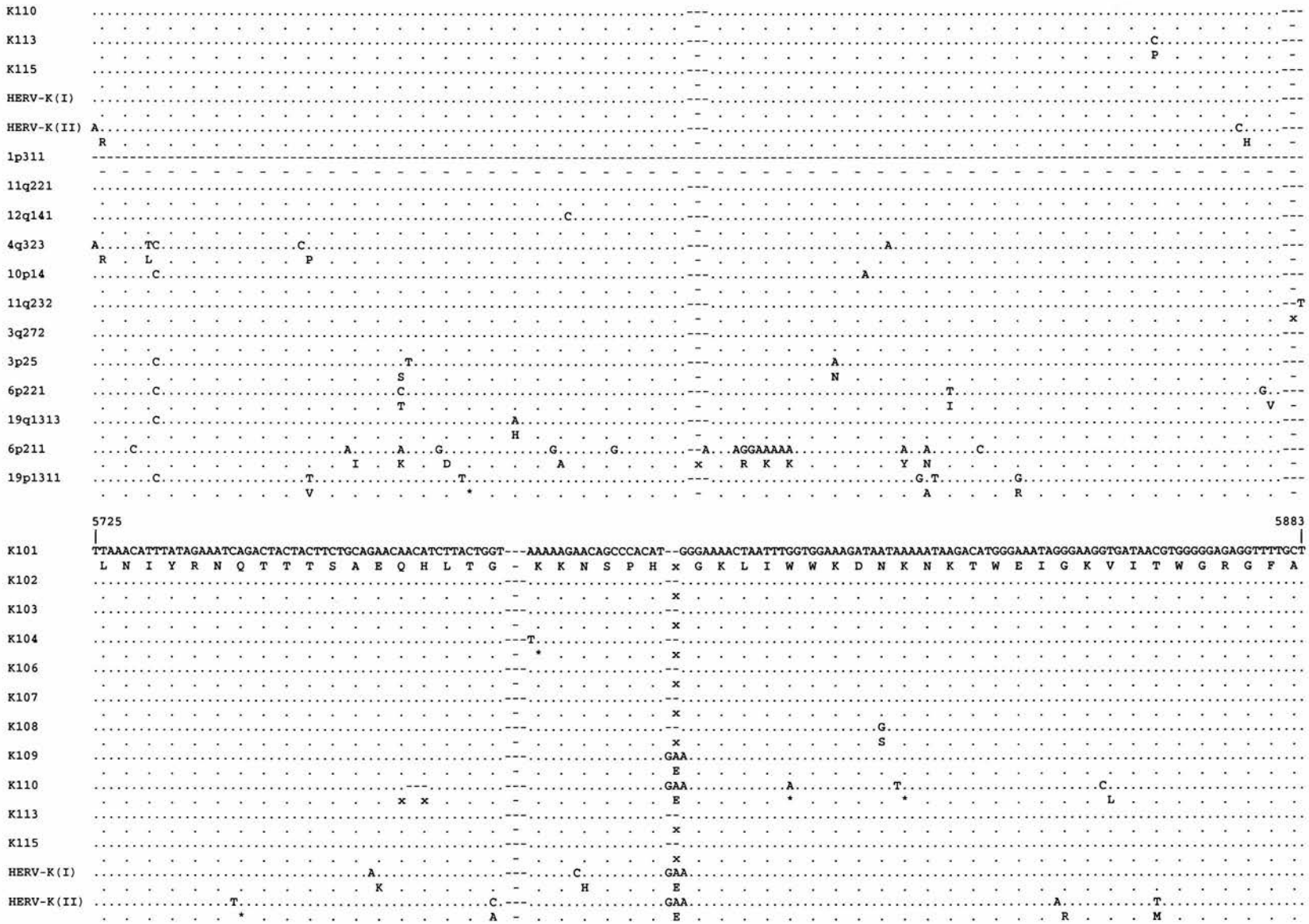

HERV-K(I)
HERV-K(II)
1p311
11q221
12q141
4q323
10p14
11q232
3q272
3p25
6p221
19q1313
6p211
19p1311
4771
4929
K101
GCAGCTTACACAGGGCCGAAAGAACGAGTAATCAAAACTCCATATCAATCGGCTCAAAGAGCAGAGTTGGTTGCAGTCATTACAGTGTTACAAGATTTTGACCAACCTATCAATATTATATCAGATTCTGCCTATGTAGTACAGGCTACAAGGGATGTT
A A Y T G P K E R V I K T P Y Q S A Q R A E L V A V I T V L Q D F D Q P I N I I S D S A Y V V Q A T R D V
K102
K103
K104
K106
K107
K108
K109
K110
K113
K115
HERV-K(I)
HERV-K(II)
1p311
11q221
12q141

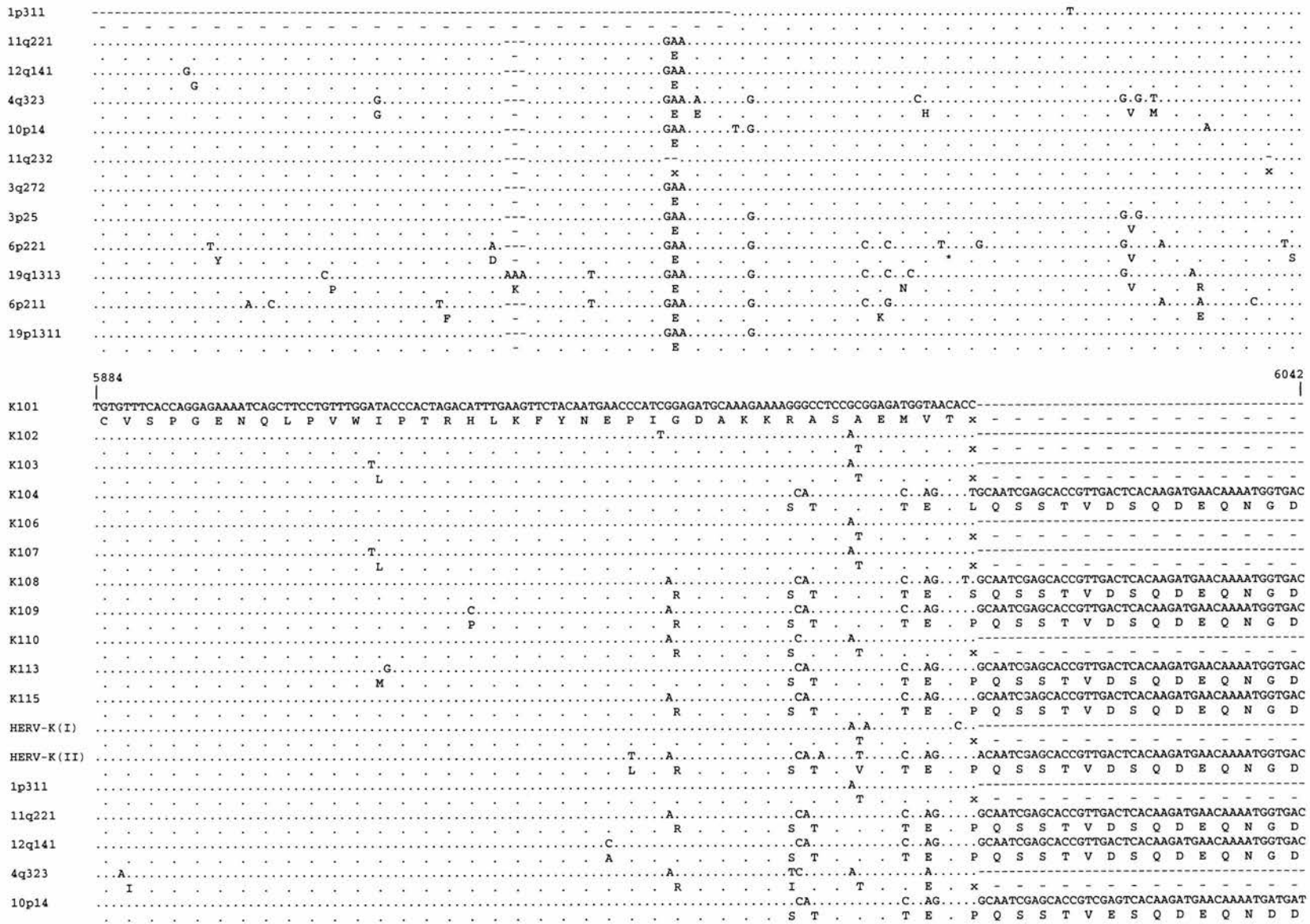

4q323
10p14
11q232
3q272
3p25
6p221
19q1313
6p211
19p1311
4930
5088
K101
E T A L I K Y S M D D Q L N Q L F N L L Q Q T V R K R N F P F Y I T H I R A H T N L P G P L T K A N E Q A
K102
K103
K104
K106
K107
K108
K109
K110
K113
K115
HERV-K(I)
HERV-K(II)
1p311
11q221
12q141
4q323
10p14
11q232
3q272
3p25


6p221
19q1313
6p211
19p1311
5089
5247
K101
D L L V S S A L I K A Q E L H A L T H V N A A G L K N K F D V T W K Q A K D I V Q H C T Q C Q V L H L P T
K102
K103
K104
K106
K107
K108
K109
K110
K113
K115
HERV-K(I)
HERV-K(II)
1p311
11q221
12q141
4q323
10p14
11q232
3q272
3p25
6p221
19q1313
6p211
19p1311

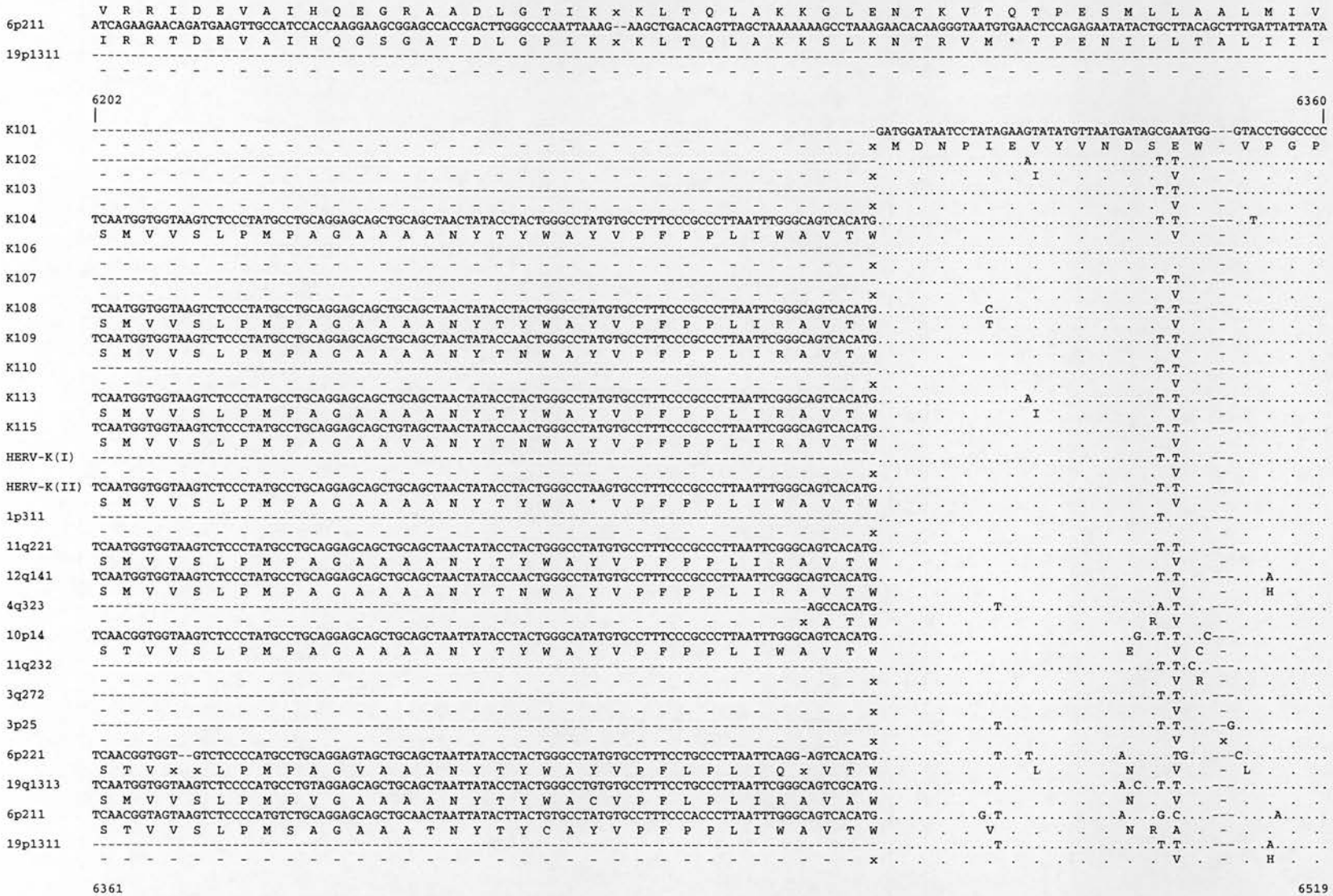

5248
5406
K101
CAAGAGGCAGGAGTTAATCCCAGAGGTCTGTGTCCTAATGCATTATGGCAAATGGATGTCACGCATGTACCTTCATTTGGA---AGATTATCATATGTTCATGTAACAGTTGATACTTATTCA------CATTTCATATGGGCAACTTGCCAAACAGGA
Q E A G V N P R G L C P N A L W Q M D V T H V P S F G - R L S Y V H V T V D T Y S - - H F I W A T C Q T G
K102
K103
K104
K106
K107
K108
K109
K110
K113
K115
HERV-K(I)
HERV-K(II)
1p311
11q221
12q141
4q323
10p14
11q232
3q272
3p25
6p221
19q1313
6p211
19p1311
5407
5565
K101
GAAAGTACTTCCCATGTTAAA---AAACATTTATTGTCTTGTTTTGCTGTAATGGGAGTTCCAGAAAAAATCAAAACTGACAATGGACCAGGATATTGTAGTAAAGCTTTCCAAAAA---------TTCTTAAGTCAGTGGAAAATTTCACATACAACA
E S T S H V K - K H L L S C F A V M G V P E K I K T D N G P G Y C S K A F Q K - - - F L S Q W K I S H T T
K102
K103
K104

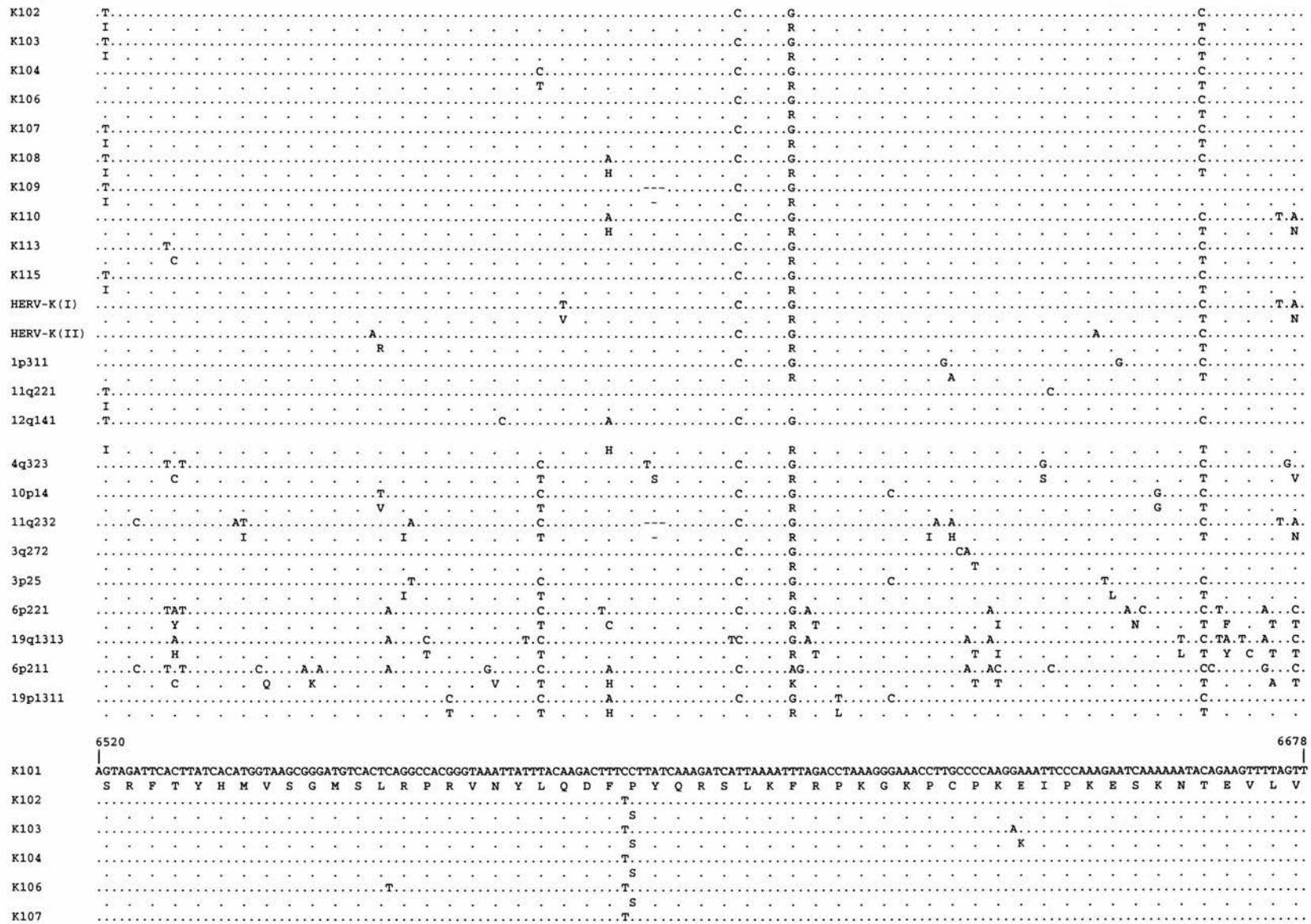

K106
K107
K108
K109
K110
K113
K115
HERV-K(I)
HERV-K(II)
1p311
11q221
12q141
4q323
10p14
11q232
3q272
3p25
6p221
19q1313
6p211
19p1311
5566
5724
K101
GGAATTCCTTATAATTCCCAAGGACAGGCCATAGTTGAAAGAACTAATAGAACACTCAAAACTCAATTAGTTAAACAA---AAAGAAGGGGGGAGACAGTAAGGAGTGTACCACTCCTCAGATGCAACTTAATCTAGCACTCTATACTTTAAATTTT---
G I P Y N S Q G Q A I V E R T N R T L K T Q L V K Q - K E G G D S K E C T T P Q M Q L N L A L Y T L N F -
K102
K103
K104
K106
K107
K108
K109

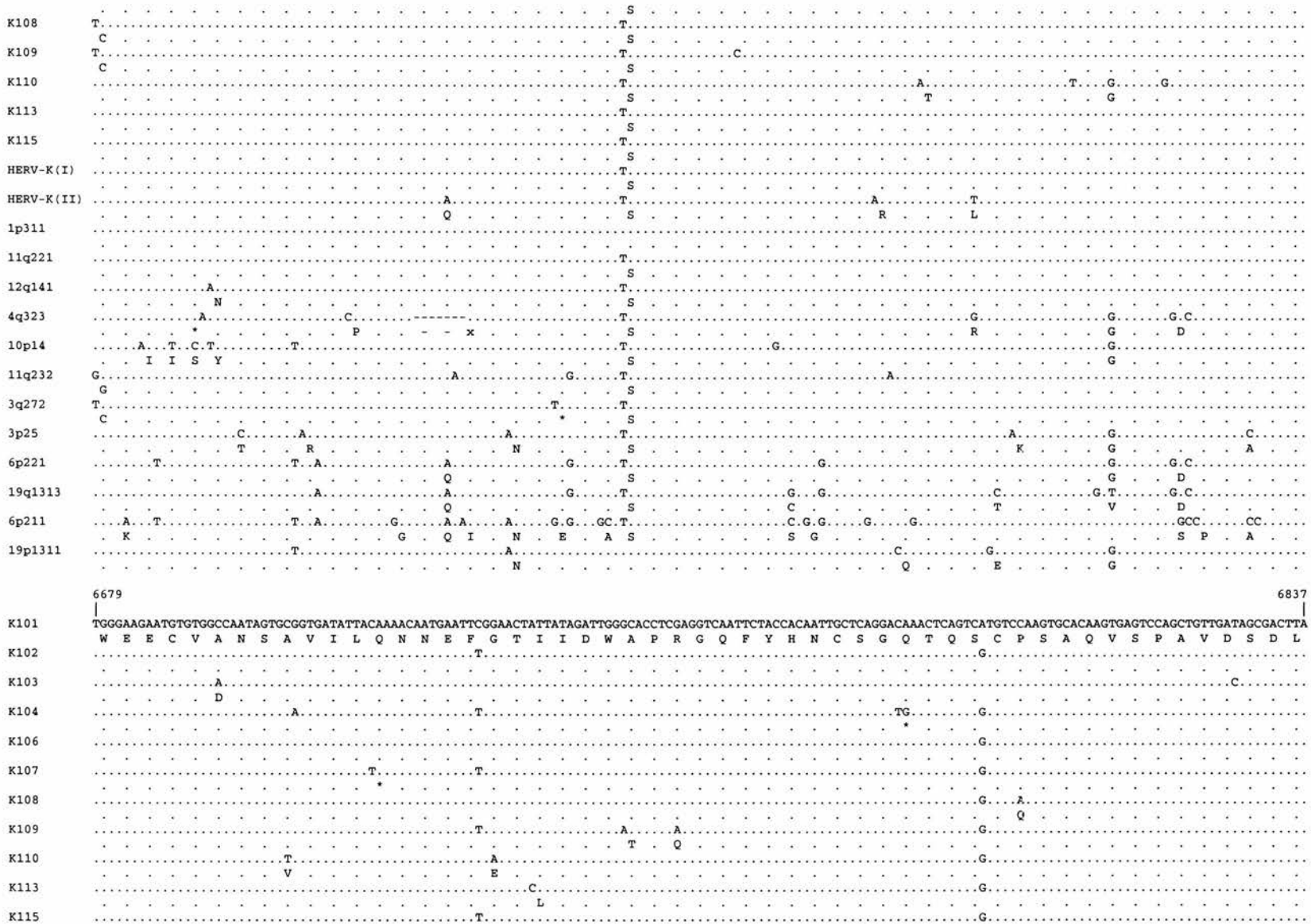

K110
K113
K115
HERV-K(I)
HERV-K(II)
1p311
11q221
12q141
4q323
10p14
11q232
3q272
3p25
6p221
19q1313
6p211
19p1311
5725
5883
TTAAACATTTATAGAAATCAGACTACTACTTCTGCAGAACAACATCTTACTGGT---AAAAAGAACAGCCCACAT--GGGAAAACTAATTTGGTGGAAAGATAATAAAAATAAGACATGGGAAATAGGGAAGGTGATAACGTGGGGGAGAGGTTTTGCT
L N I Y R N Q T T T S A E Q H L T G - K K N S P H x G K L I W W K D N K N K T W E I G K V I T W G R G F A
K102
K103
K104
K106
K107
K108
K109
K110
K113
K115
HERV-K(I)
HERV-K(II)

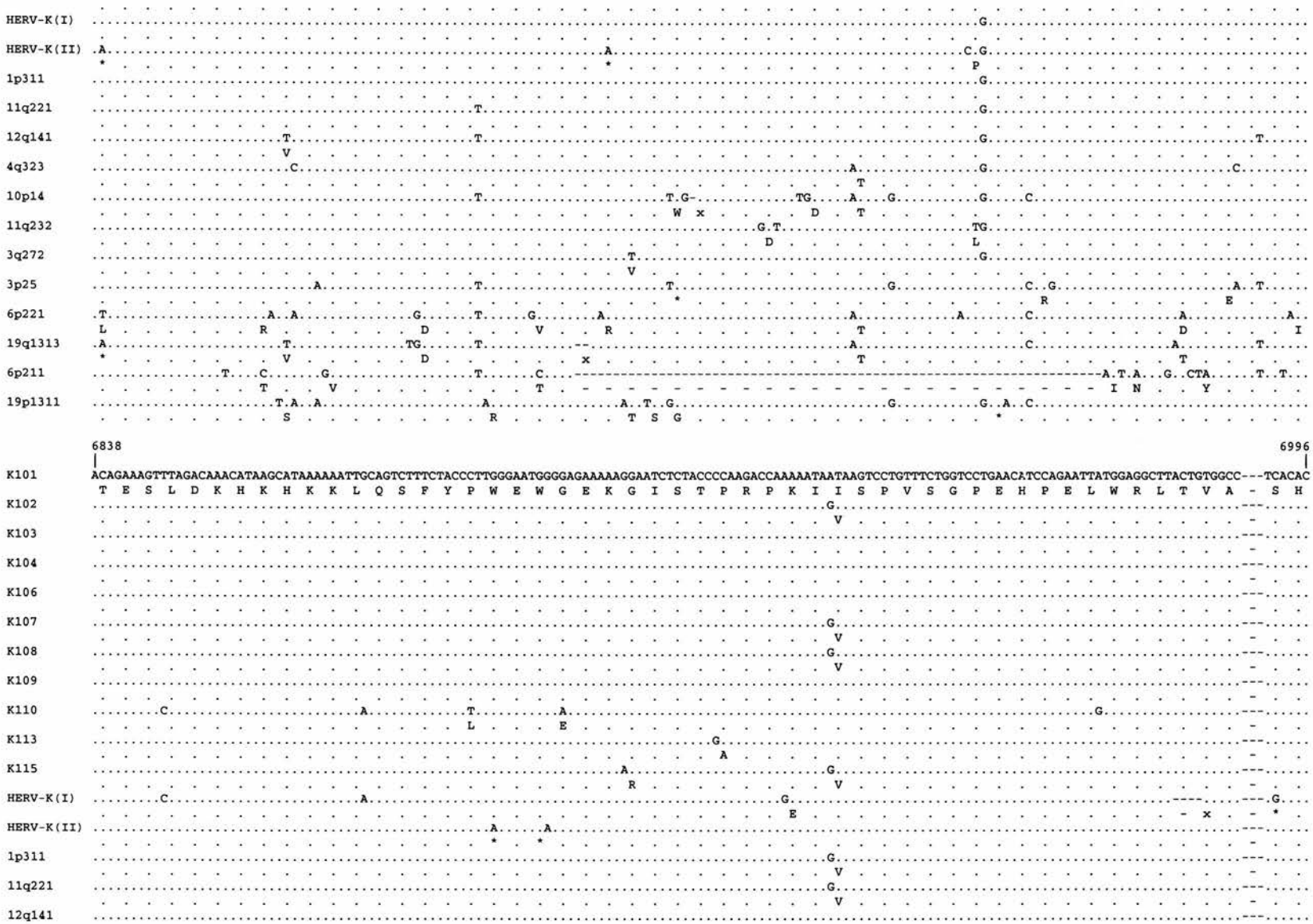

1p311
11q221
12q141
4q323
10p14
11q232
3q272
3p25
6p221
19q1313
6p211
19p1311
5884
6042
K101
K102
K103
K104
K106
K107
K108
K109
K110
K113
K115
HERV-K(I)
HERV-K(II)
1p311
11q221
12q141
4q323
10p14

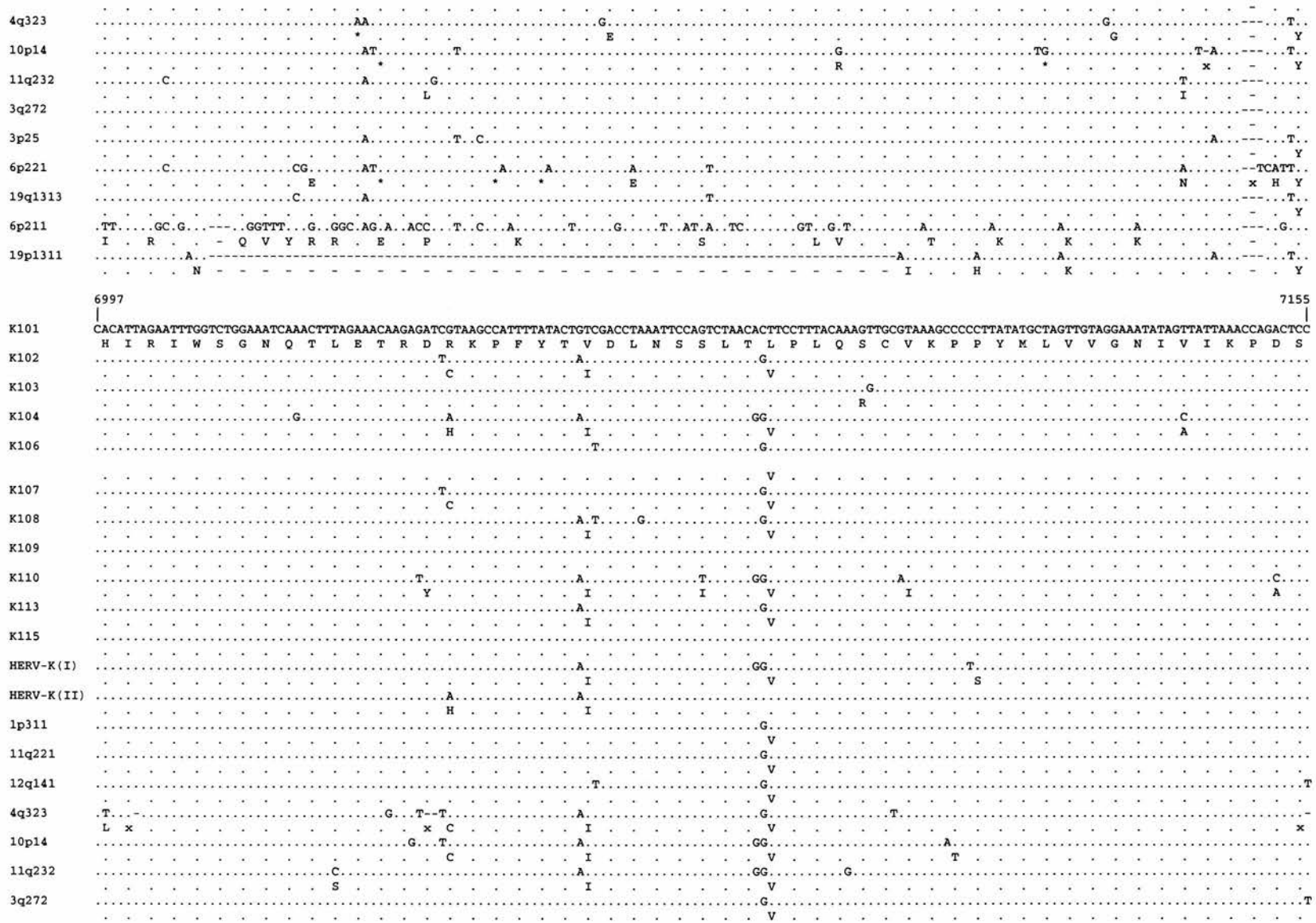

11q232
3q272
3p25
6p221
19q1313
6p211
19p1311
6043
6201
K101
K102
K103
K104
K106
K107
K108
K109
K110
K113
K115
HERV-K(I)
HERV-K(II)
1p311
11q221
12q141
4q323
10p14
11q232
3q272
3p25
6p221
19q1313

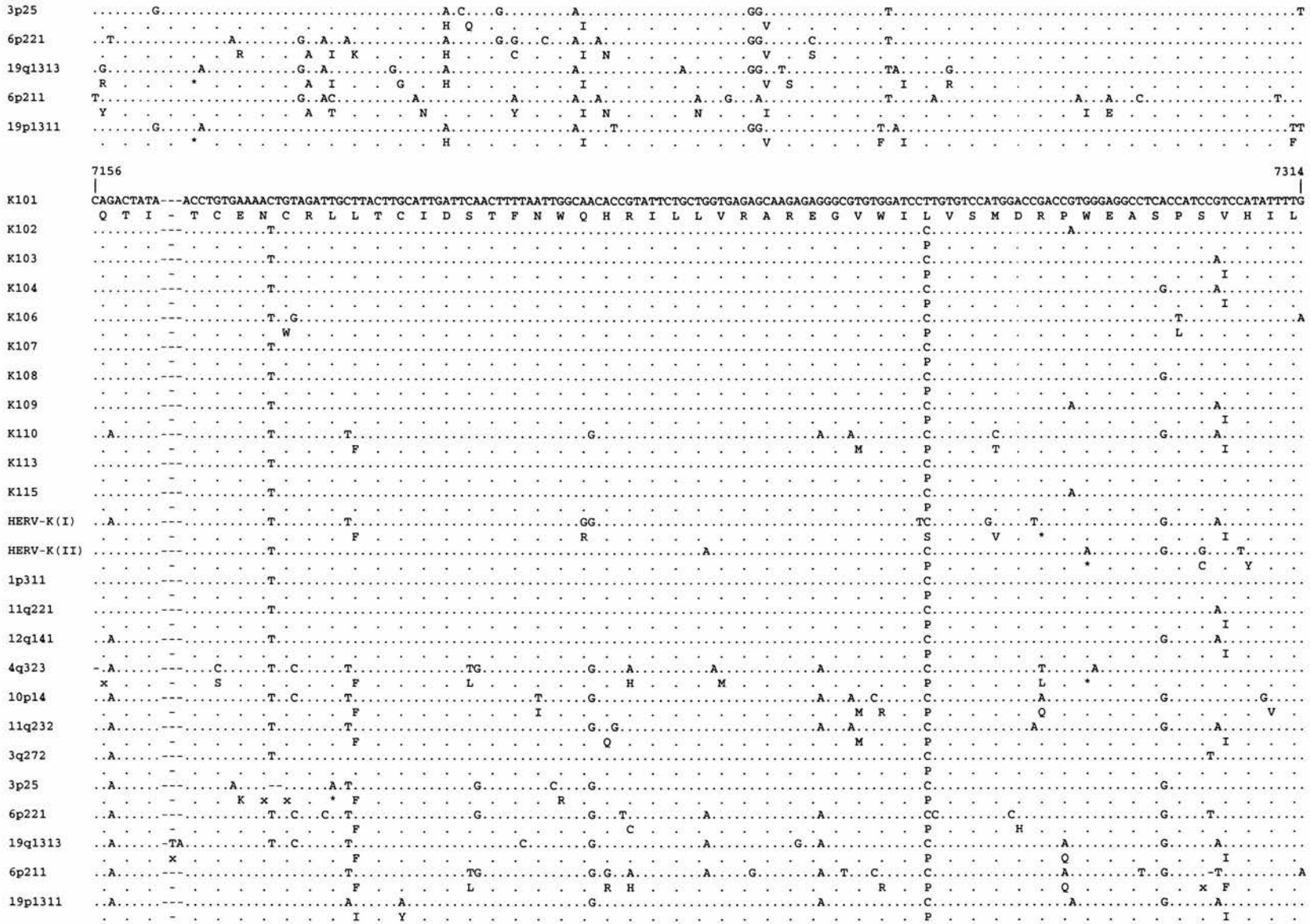


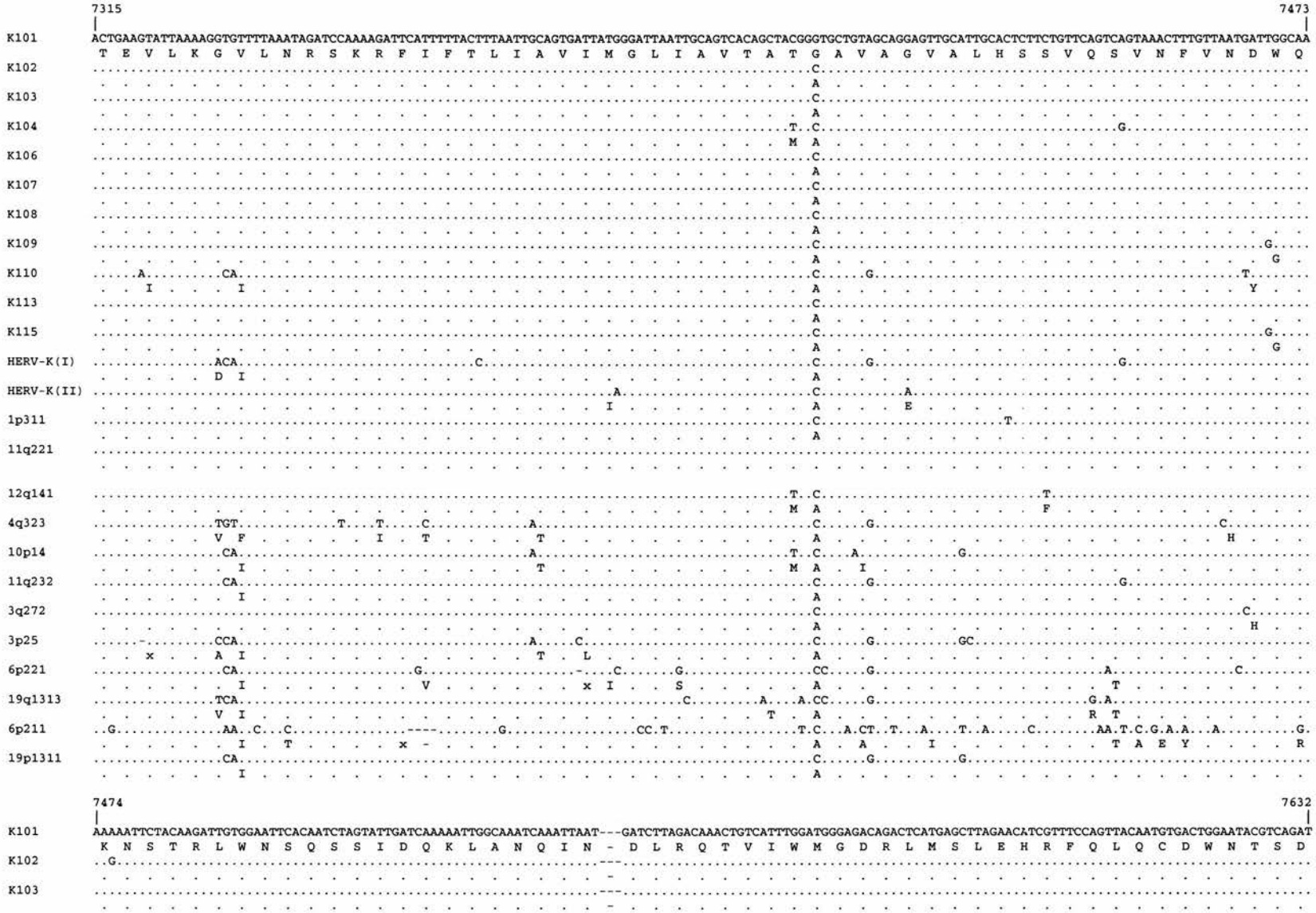

K102
K103
K104
K106
K107
K108
K109
K110
K113
K115
HERV-K(I)
HERV-K(II)
1p311
11q221
12q141
4q323
10p14
11q232
3q272
3p25
6p221
19q1313
6p211
19p1311
6520
6678
K101
S R F T Y H M V S G M S L R P R V N Y L Q D F P Y Q R S L K F R P K G K P C P K E I P K E S K N T E V L V
K102
K103
K104
K106
K107

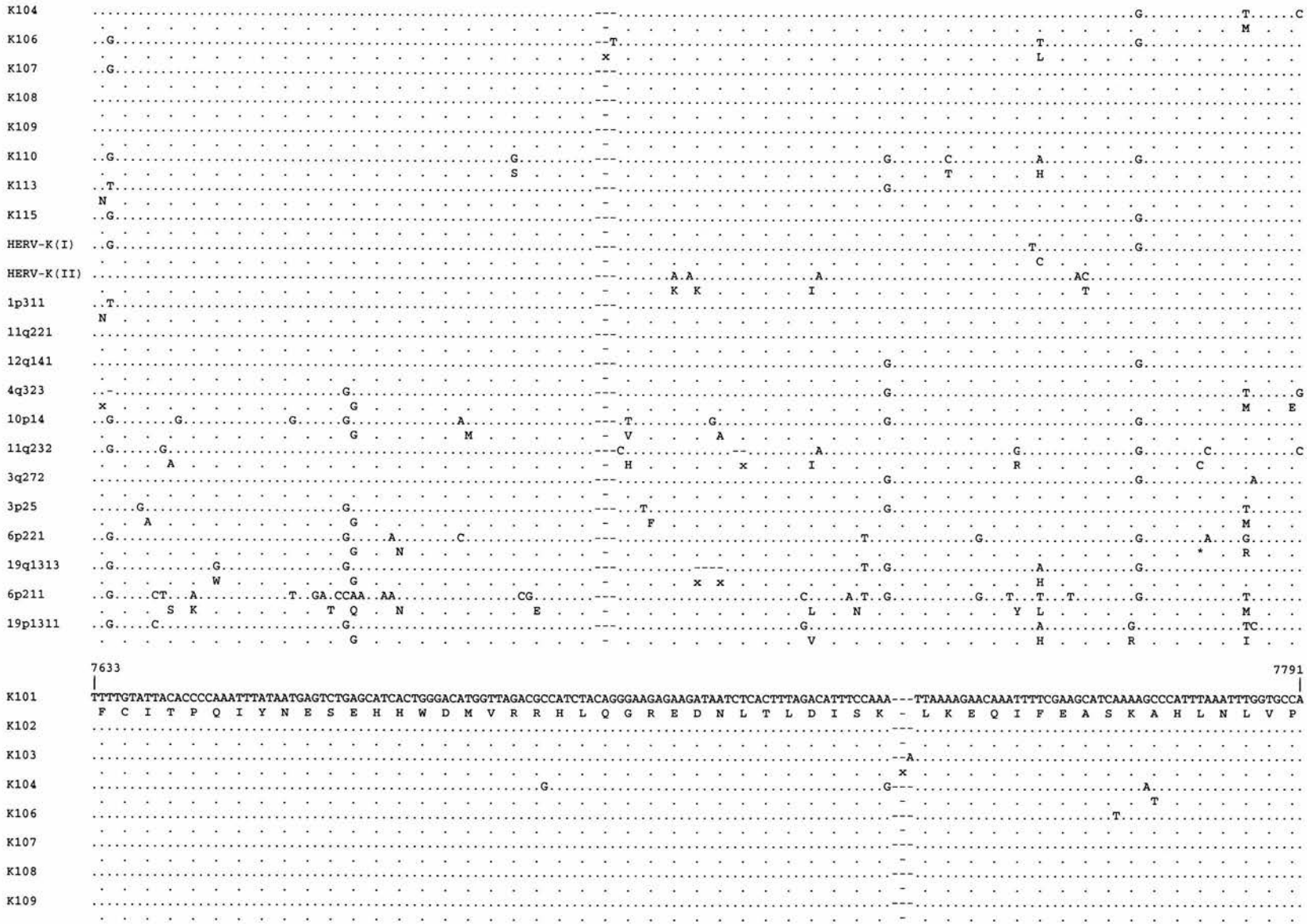

K108
K109
K110
K113
K115
HERV-K(I)
HERV-K(II)
1p311
11q221
12q141
4q323
10p14
11q232
3q272
3p25
6p221
19q1313
6p211
19p1311
6679
6837
K101
K102
K103
K104
K106
K107
K108
K109
K110
K113
K115

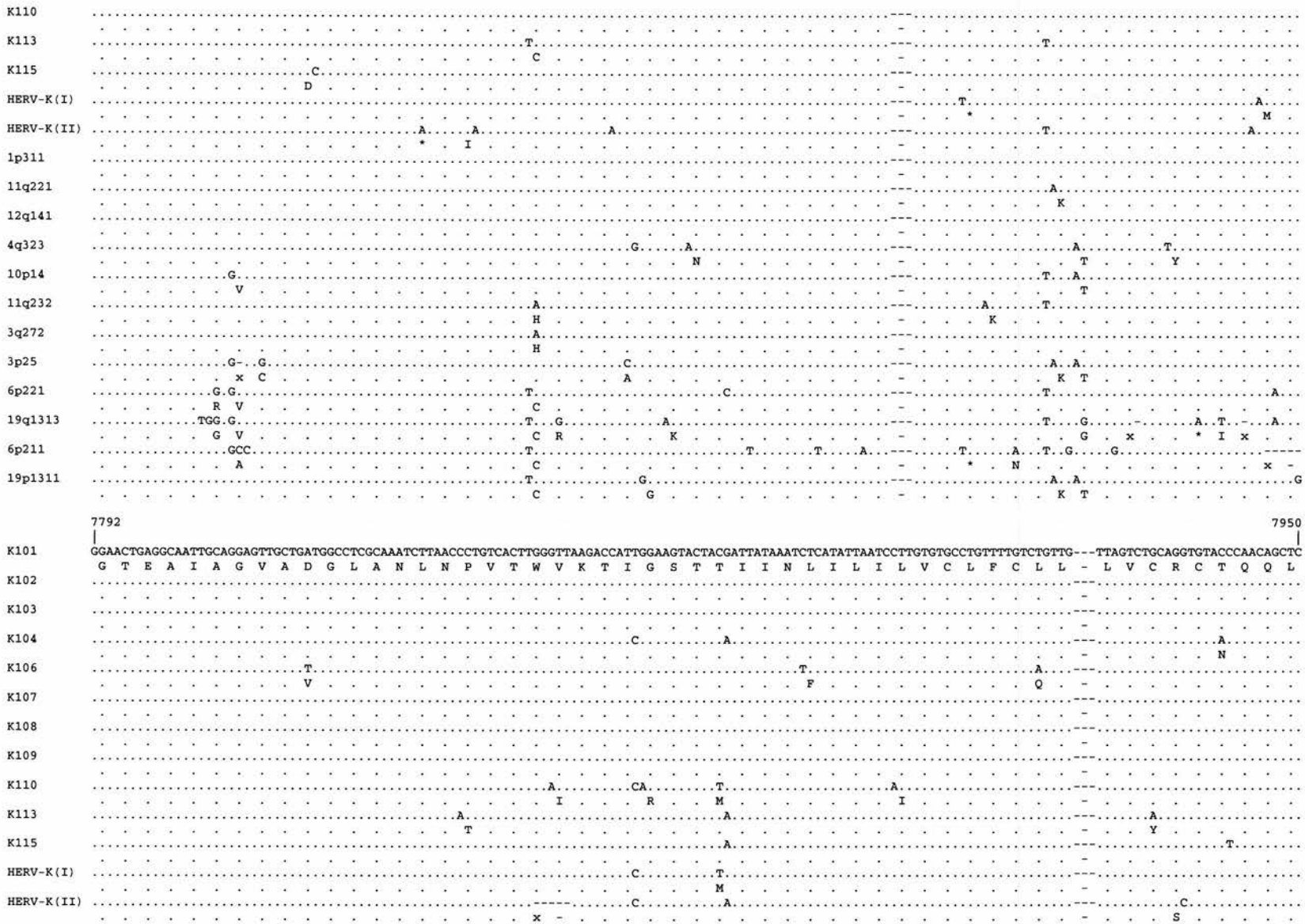

HERV-K(I)
HERV-K(II)
1p311
11q221
12q141
4q323
10p14
11q232
3q272
3p25
6p221
19q1313
6p211
19p1311
6838
6996
K101
ACAGAAAGTTTAGACAAACATAAGCATAAAAAATTGCAGTCTTTCTACCCTTGGGAATGGGGAGAAAAAGGAATCTCTACCCCAAGACCAAAAATAATAAGTCCTGTTTCTGGTCCTGAACATCCAGAATTATGGAGGCTTACTGTGGCC---TCACAC
T E S L D K H K H K K L Q S F Y P W E W G E K G I S T P R P K I I S P V S G P E H P E L W R L T V A - S H
K102
K103
K104
K106
K107
K108
K109
K110
K113
K115
HERV-K(I)
HERV-K(II)
1p311
11q221
12q141

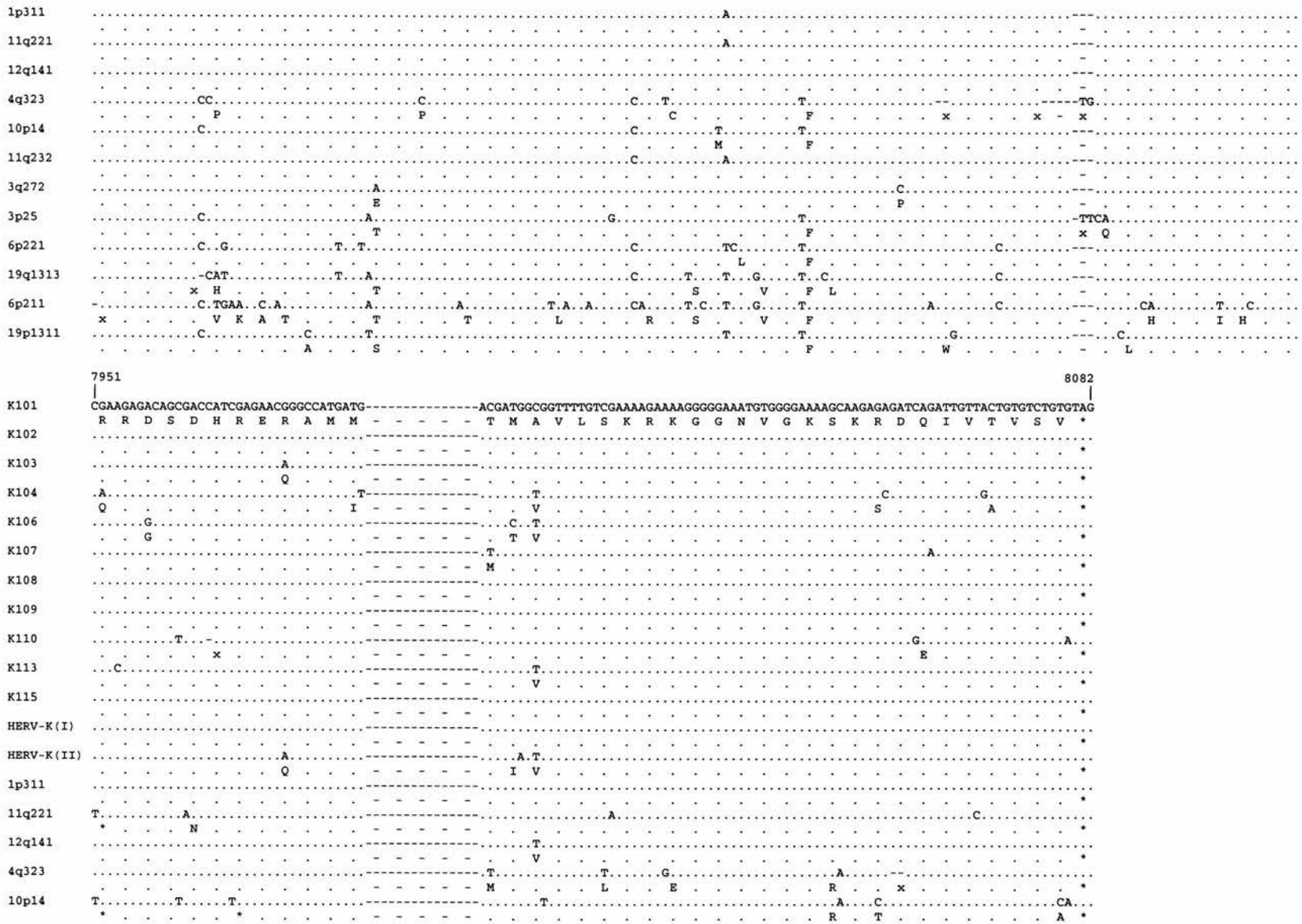

4q323
10p14
11q232
3q272
3p25
6p221
19q1313
6p211
19p1311
6997
7155
K101
CACATTAGAATTTGGTCTGGAAATCAAACTTTAGAAACAAGAGATCGTAAGCCATTTTATACTGTCGACCTAAATTCCAGTCTAACACTTCCTTTACAAAGTTGCGTAAAGCCCCCTTATATGCTAGTTGTAGGAAATATAGTTATTAAACCAGACTCC
H I R I W S G N Q T L E T R D R K P F Y T V D L N S S L T L P L Q S C V K P P Y M L V V G N I V I K P D S
K102
K103
K104
K106
K107
K108
K109
K110
K113
K115
HERV-K(I)
HERV-K(II)
1p311
11q221
12q141
4q323
10p14
11q232
3q272

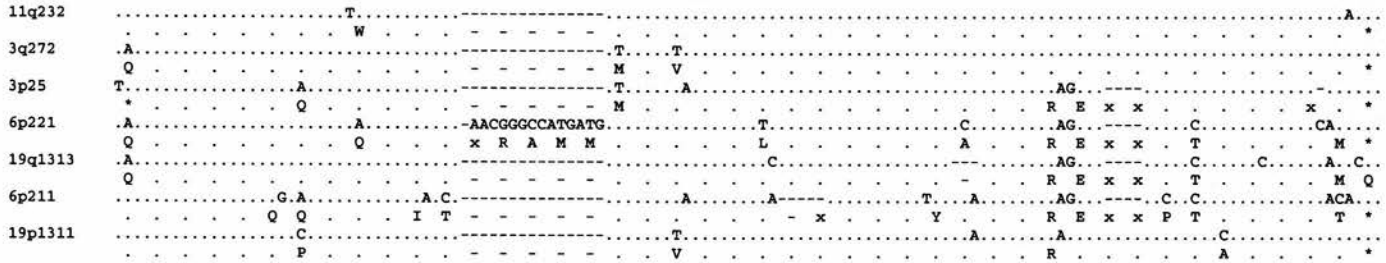

7156
7314
K101 CAGACTATA---ACCTGTGAAAACTGTAGATTGCTTACTTGCATTGATTCAACTTTTAATTGGCAACACCGTATTCTGCTGGTGAGAGCAAGAGAGGGCGTGTGGATCCTTGTGTCCATGGACCGACCGTGGGAGGCCTCACCATCCGTCCATATTTTG
Q T I - T C E N C R L L T C I D S T F N W Q H R I L L V R A R E G V W I L V S M D R P W E A S P S V H I L
K102
K103
K104
K106
K107
K108
K109
K110
K113
K115
HERV-K(I)
HERV-K(II)
1p311
11q221
12q141
4q323
10p14
11q232
3q272
3p25
6p221
19q1313
6p211
19p1311


7315
7473
K101
ACTGAAGTATTAAAAGGTGTTTTAAATAGATCCAAAAGATTCATTTTTACTTTAATTGCAGTGATTATGGGATTAATTGCAGTCACAGCTACGGGTGCTGTAGCAGGAGTTGCATTGCACTCTTCTGTTCAGTCAGTAAACTTTGTTAATGATTGGCAA
T E V L K G V L N R S K R F I F T L I A V I M G L I A V T A T G A V A G V A L H S S V Q S V N F V N D W Q
K102
K103
K104
K106
K107
K108
K109
K110
K113
K115
HERV-K(I)
HERV-K(II)
1p311
11q221
12q141
4q323
10p14
11q232
3q272
3p25
6p221
19q1313
6p211
19p1311
7474
7632
K101
AAAAATTCTACAAGATTGTGGAATTCACAATCTAGTATTGATCAAAAATTGGCAAATCAAATTAAT---GATCTTAGACAAACTGTCATTTGGATGGGAGACAGACTCATGAGCTTAGAACATCGTTTCCAGTTACAATGTGACTGGAATACGTCAGAT
K N S T R L W N S Q S S I D Q K L A N Q I N - D L R Q T V I W M G D R L M S L E H R F Q L Q C D W N T S D
K102
K103

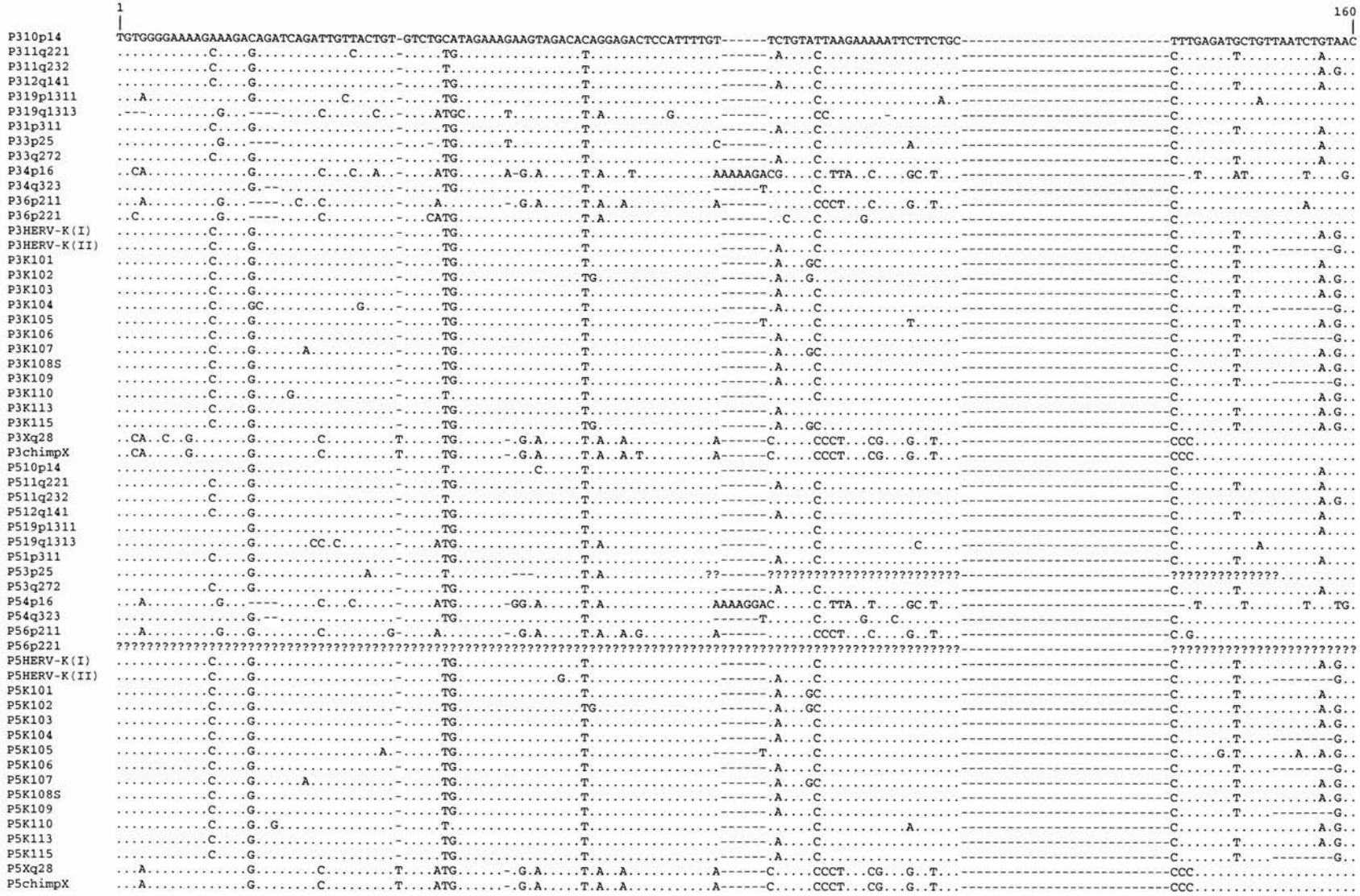

K104
K106
K107
K108
K109
K110
K113
K115
HERV-K(I)
HERV-K(II)
1p311
11q221
12q141
4q323
10p14
11q232
3q272
3p25
6p221
19q1313
6p211
19p1311
7633
7791
K101
TTTTGTATTACACCCCAAATTTATAATGAGTCTGAGCATCACTGGGACATGGTTAGACGCCATCTACAGGGAAGAGAAGATAATCTCACTTTAGACATTTCCAAA---TTAAAAGAACAAATTTTCGAAGCATCAAAAGCCCATTTAAATTTGGTGCCA
F C I T P Q I Y N E S E H H W D M V R R H L Q G R E D N L T L D I S K - L K E Q I F E A S K A H L N L V P
K102
K103
K104
K106
K107
K108
K109

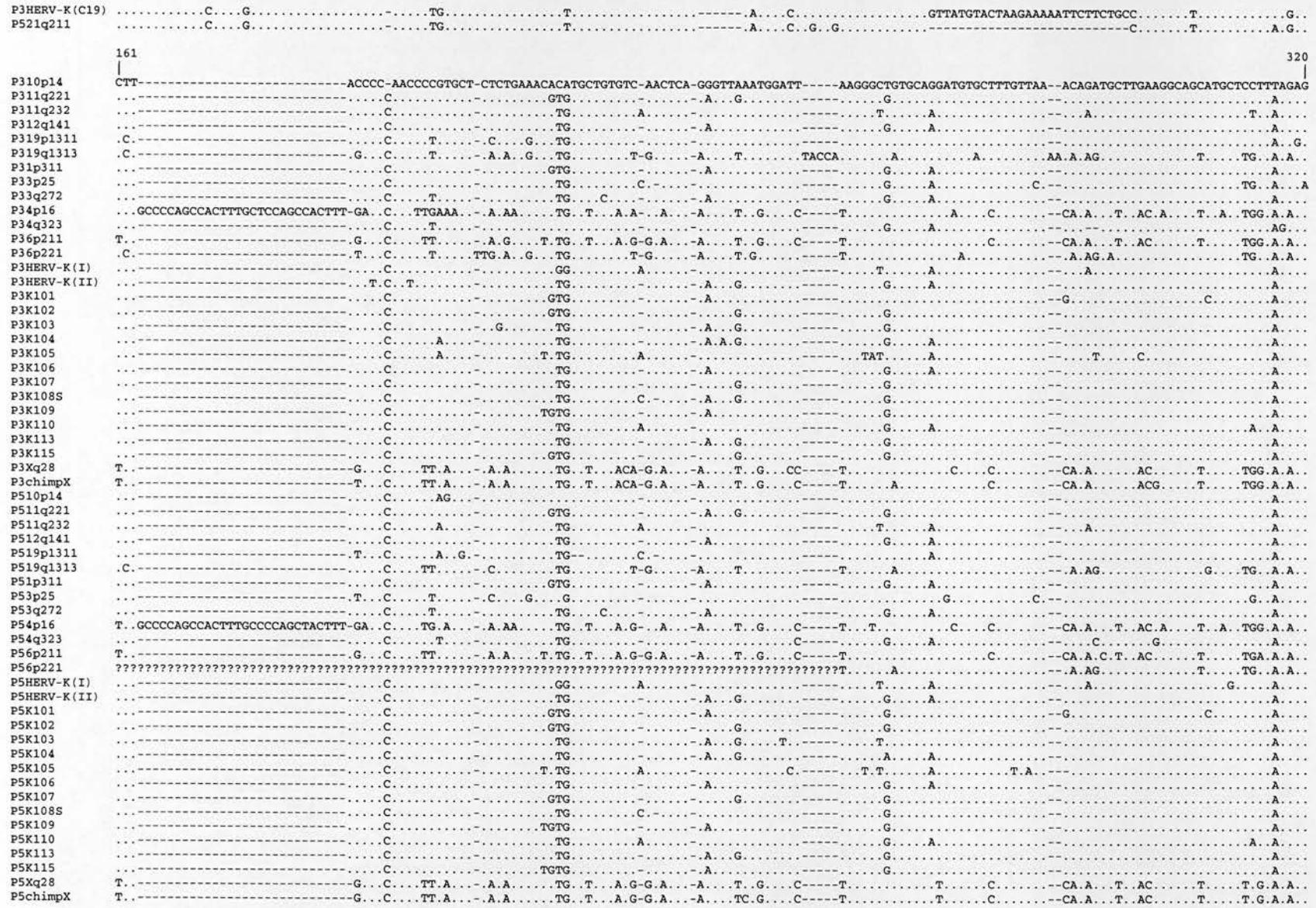

K110
K113
K115
HERV-K(I)
HERV-K(II)
1p311
11q221
12q141
4q323
10p14
11q232
3q272
3p25
6p221
19q1313
6p211
19p1311
7792
7950
K101
GGAACTGAGGCAATTGCAGGAGTTGCTGATGGCCTCGCAAATCTTAACCCTGTCACTTGGGTTAAGACCATTGGAAGTACTACGATTATAAATCTCATATTAATCCTTGTGTGCCTGTTTTGTCTGTTG---TTAGTCTGCAGGTGTACCCAACAGCTC
G T E A I A G V A D G L A N L N P V T W V K T I G S T T I I N L I L I L V C L F C L L - L V C R C T Q Q L
K102
K103
K104
K106
K107
K108
K109
K110
K113
K115
HERV-K(I)
HERV-K(II)

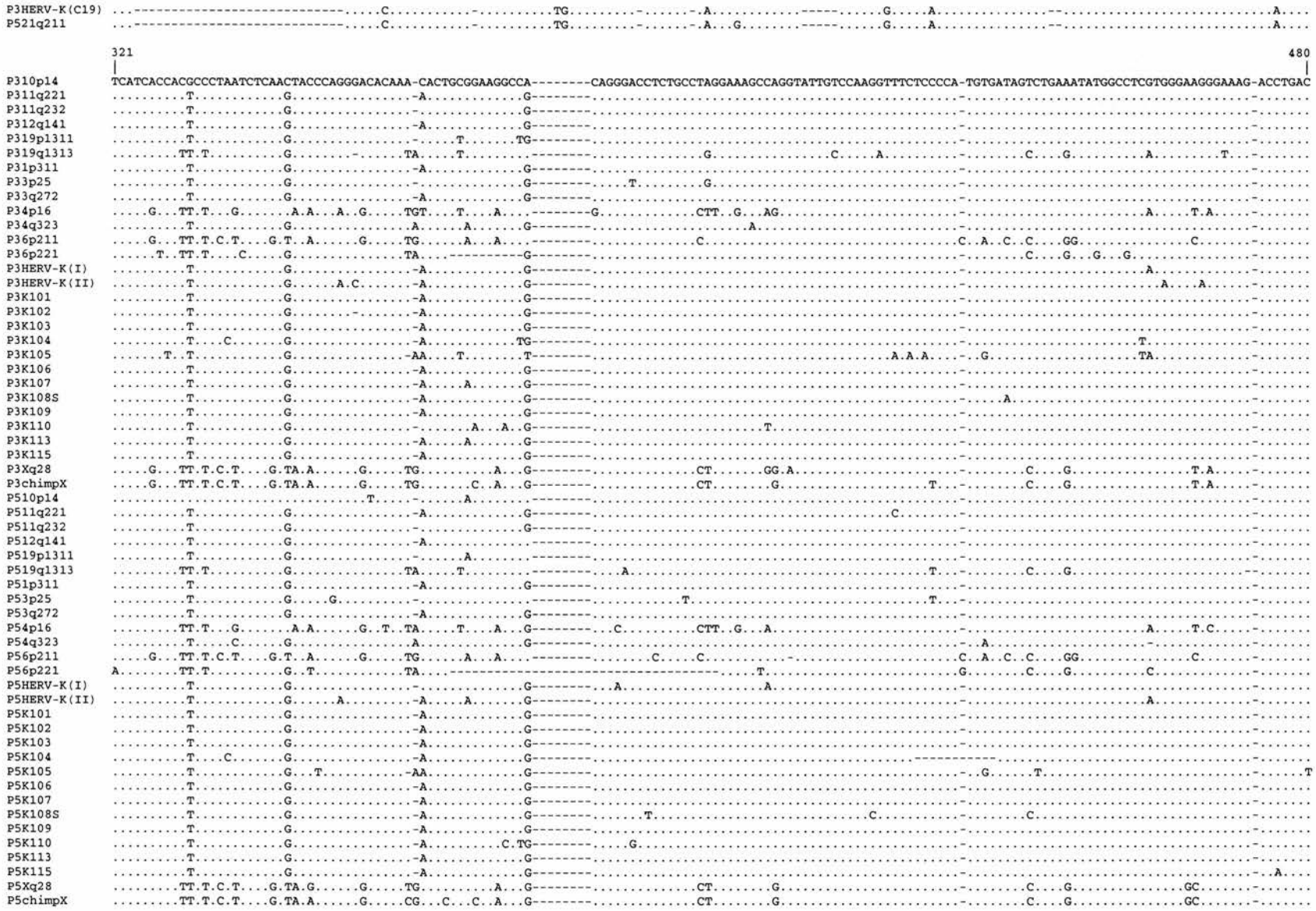

1p311
11q221
12q141
4q323
10p14
11q232
3q272
3p25
6p221
19q1313
6p211
19p1311
7951
8082
K101
CGAAGAGACAGCGACCATCGAGAACGGGCCATGATG----------ACGATGGCGGTTTTGTCGAAAAGAAAAGGGGGAAATGTGGGGAAAAGCAAGAGAGATCAGATTGTTACTGTGTCTGTGTAG
R R D S D H R E R A M M - - - - - T M A V L S K R K G G N V G K S K R D Q I V T V S V *
K102
K103
K104
K106
K107
K108
K109
K110
K113
K115
HERV-K(I)
HERV-K(II)
1p311
11q221
12q141
4q323
10p14

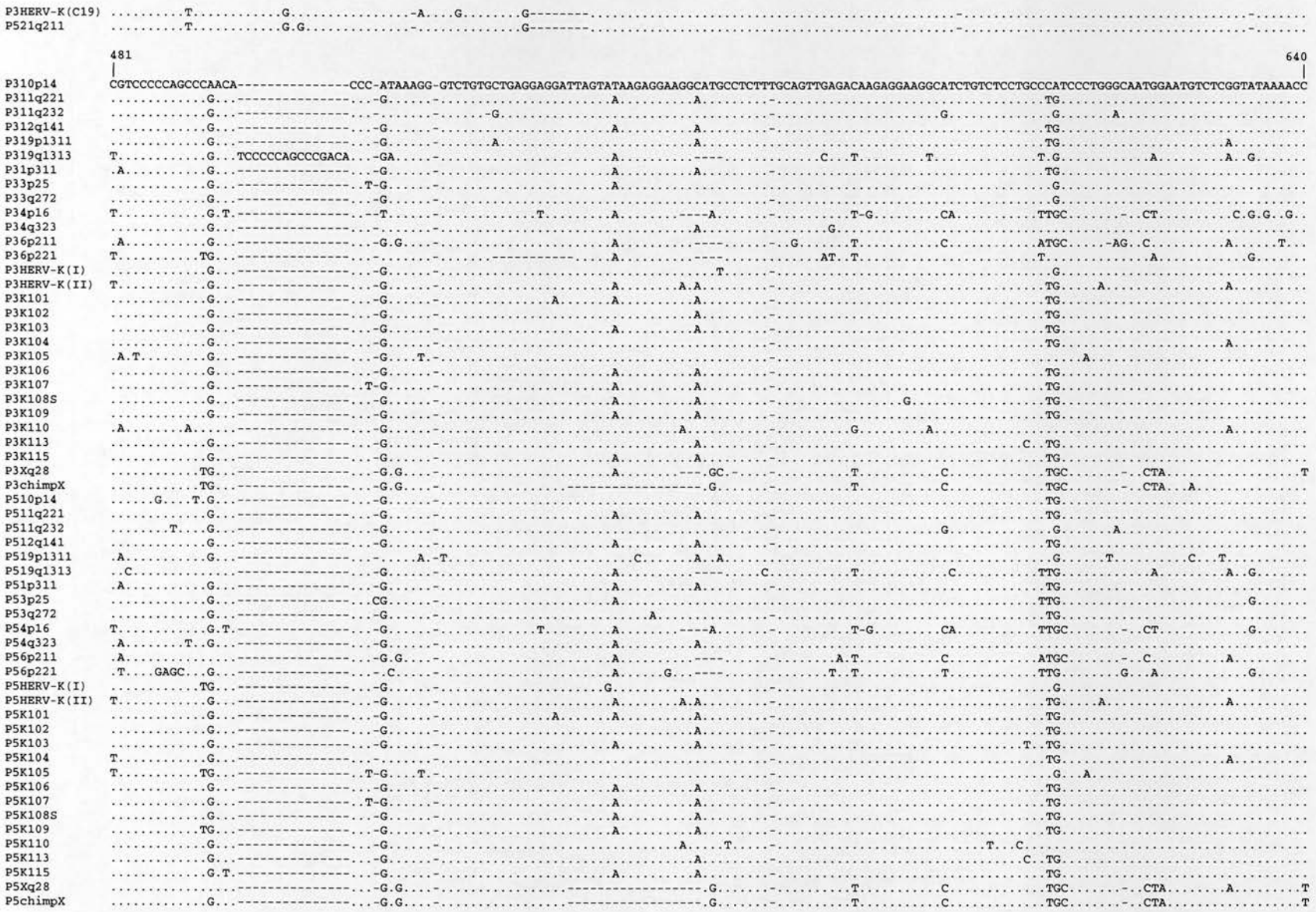

11q232
3q272
3p25
6p221
19q1313
6p211
19p1311

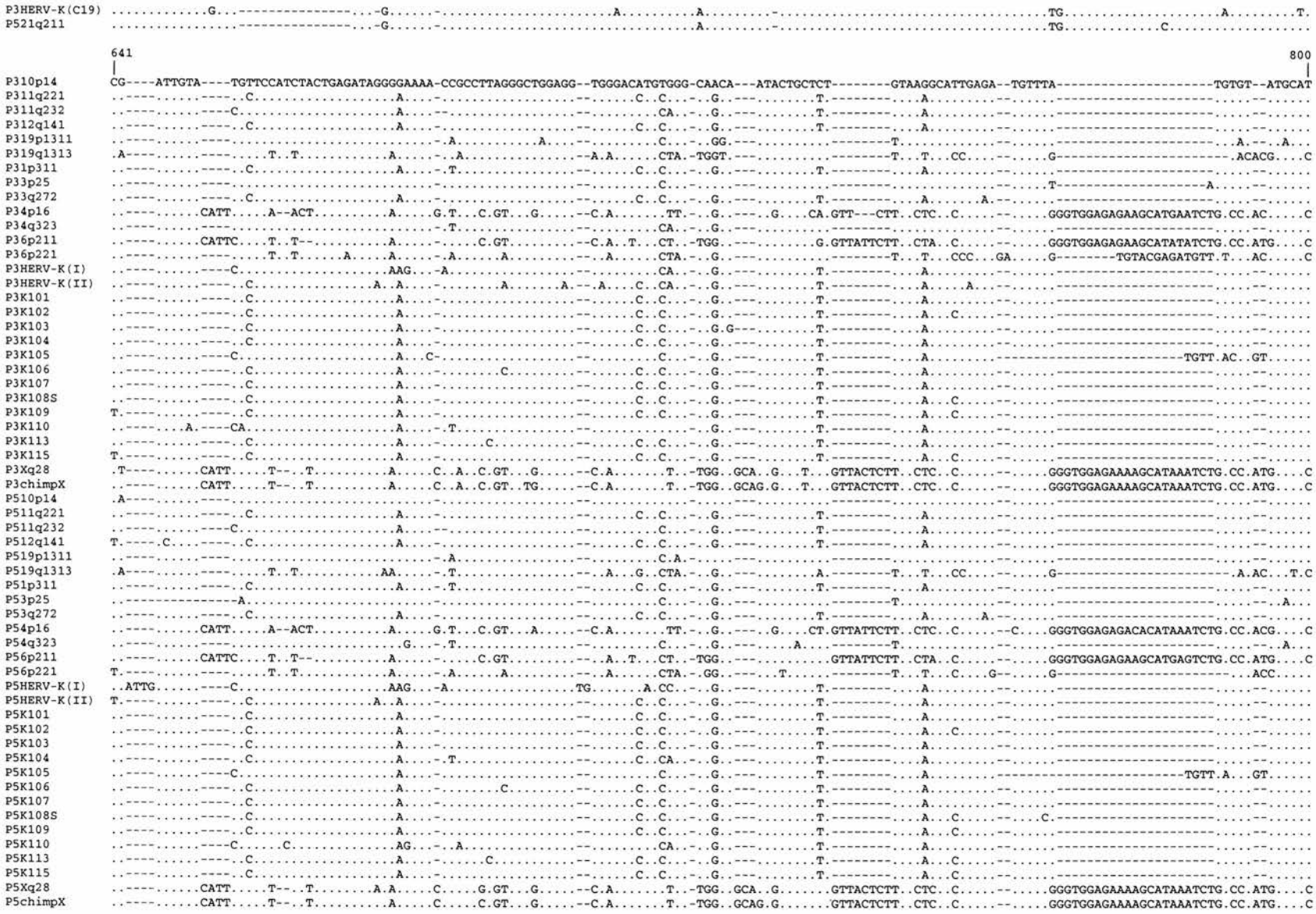
A.2 HERV-K(HML-2) Alignment of the respective LTRs of individual proviruses

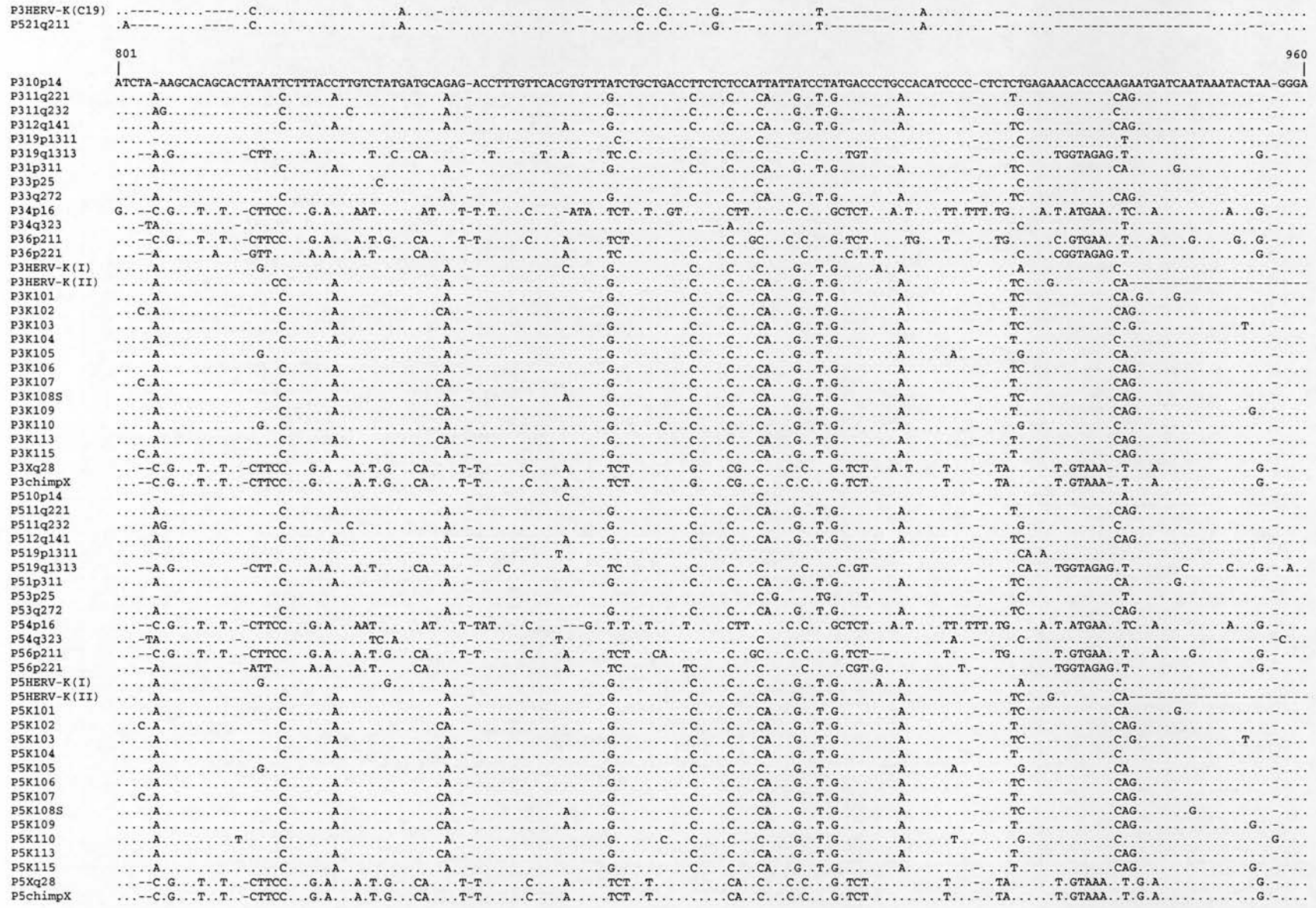


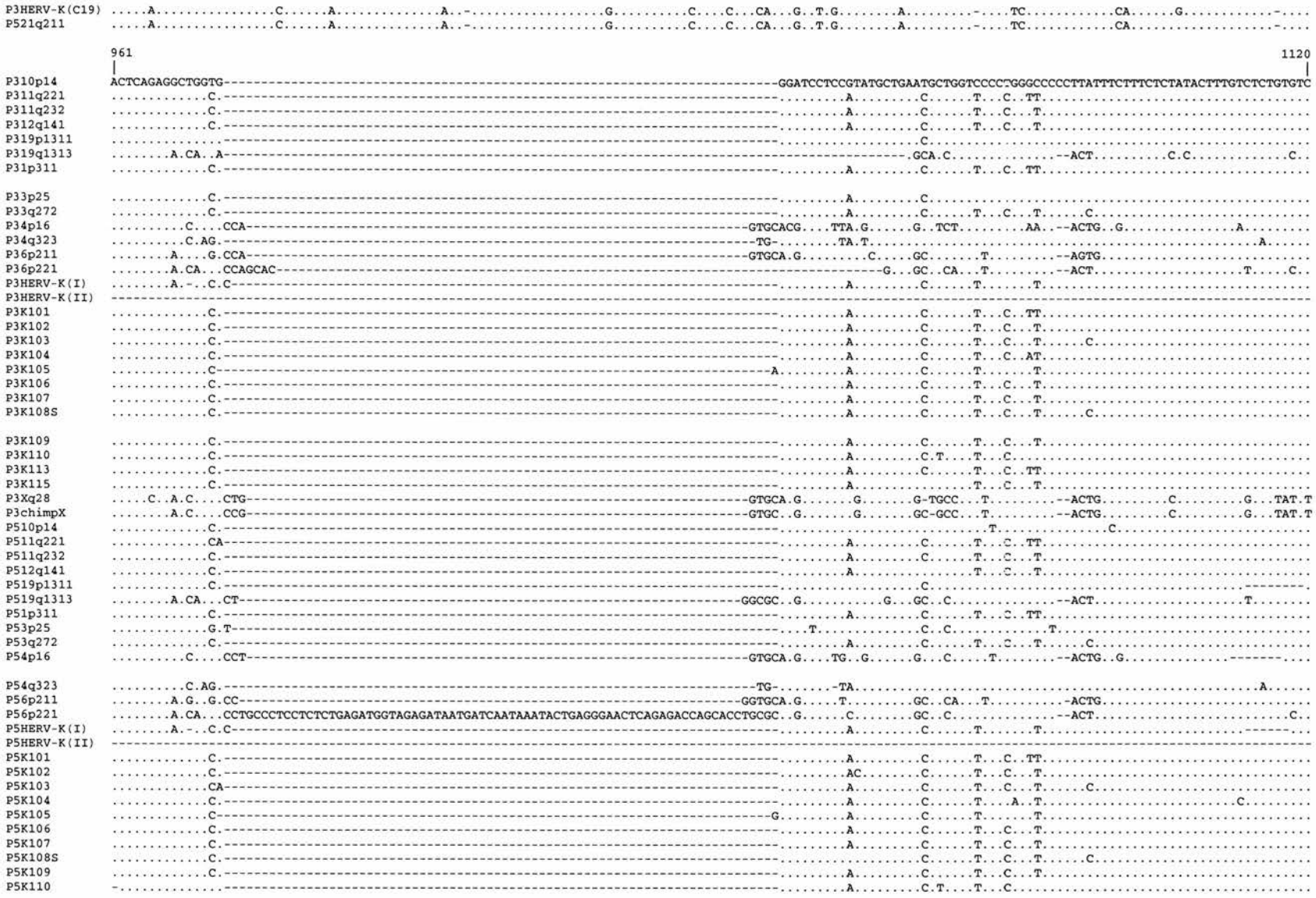

P3HERV-K(C19)
P521q211
161
320
P310p14
P311q221
P311q232
P312q141
P319p1311
P319q1313
P31p311
P33p25
P33q272
P34p16
P34q323
P36p211
P36p221
P3HERV-K(I)
P3HERV-K(II)
P3K101
P3K102
P3K103
P3K104
P3K105
P3K106
P3K107
P3K108S
P3K109
P3K110
P3K113
P3K115
P3Xq28
P3chimpX
P510p14
P511q221
P511q232
P512q141
P519p1311
P519q1313
P51p311
P53p25
P53q272
P54p16
P54q323
P56p211
P56p221
P5HERV-K(I)
P5HERV-K(II)
P5K101
P5K102
P5K103
P5K104
P5K105
P5K106
P5K107
P5K108S
P5K109
P5K110
P5K113
P5K115
P5Xq28
P5chimpX

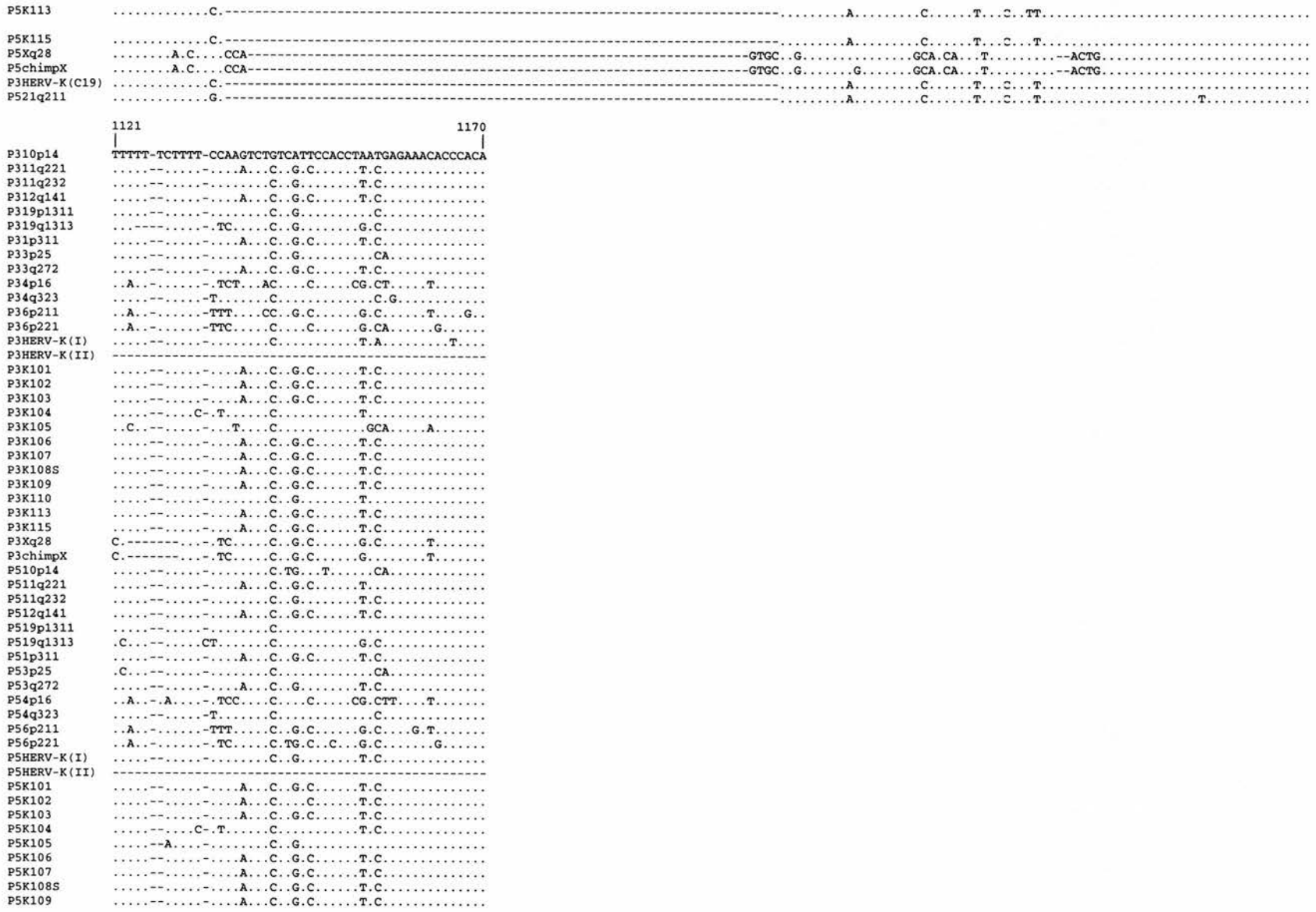

P3HERV-K(C19)
P521q211
321
480
P310p14
P311q221
P311q232
P312q141
P319p1311
P319q1313
P31p311
P33p25
P33q272
P34p16
P34q323
P36p211
P36p221
P3HERV-K(I)
P3HERV-K(II)
P3K101
P3K102
P3K103
P3K104
P3K105
P3K106
P3K107
P3K108S
P3K109
P3K110
P3K113
P3K115
P3Xq28
P3chimpX
P510p14
P511q221
P511q232
P512q141
P519p1311
P519q1313
P51p311
P53p25
P53q272
P54p16
P54q323
P56p211
P56p221
P5HERV-K(I)
P5HERV-K(II)
P5K101
P5K102
P5K103
P5K104
P5K105
P5K106
P5K107
P5K108S
P5K109
P5K110
P5K113
P5K115
P5Xq28
P5chimpX


P3HERV-K(C19)
P521q211
481
640
P310p14
P311q221
P311q232
P312q141
P319p1311
P319q1313
P31p311
P33p25
P33q272
P34p16
P34q323
P36p211
P36p221
P3HERV-K(I)
P3HERV-K(II)
P3K101
P3K102
P3K103
P3K104
P3K105
P3K106
P3K107
P3K108S
P3K109
P3K110
P3K113
P3K115
P3Xq28
P3chimpX
P510p14
P511q221
P511q232
P512q141
P519p1311
P519q1313
P51p311
P53p25
P53q272
P54p16
P54q323
P56p211
P56p221
P5HERV-K(I)
P5HERV-K(II)
P5K101
P5K102
P5K103
P5K104
P5K105
P5K106
P5K107
P5K108S
P5K109
P5K110
P5K113
P5K115
P5Xq28
P5chimpX


641
800
P3HERV-K(C19)
P521q211
P310p14
P311q221
P311q232
P312q141
P319p1311
P319q1313
P31p311
P33p25
P33q272
P34p16
P34q323
P36p211
P36p221
P3HERV-K(I)
P3HERV-K(II)
P3K101
P3K102
P3K103
P3K104
P3K105
P3K106
P3K107
P3K108S
P3K109
P3K110
P3K113
P3K115
P3Xq28
P3chimpX
P510p14
P511q221
P511q232
P512q141
P519p1311
P519q1313
P51p311
P53p25
P53q272
P54p16
P54q323
P56p211
P56p221
P5HERV-K(I)
P5HERV-K(II)
P5K101
P5K102
P5K103
P5K104
P5K105
P5K106
P5K107
P5K108S
P5K109
P5K110
P5K113
P5K115
P5Xq28
P5chimpX

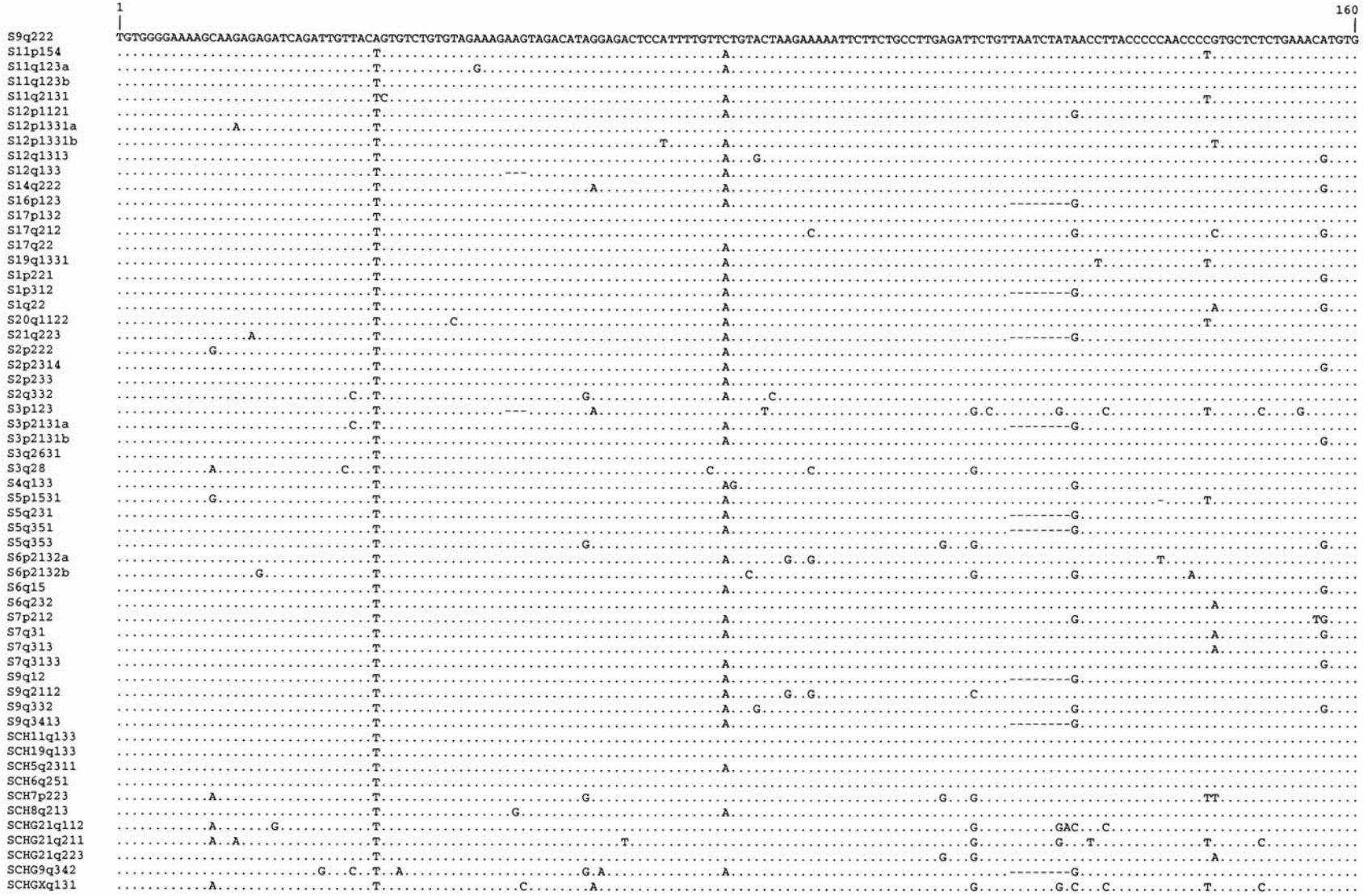

P3HERV-K(C19)
P521q211
801
960
P310p14 ATCTA-AAGCACAGCACTTAATTCTTTACCTTGTCTATGATGCAGAG-ACCTTTGTTCACGTGTTTATCTGCTGACCTTCTCTCCATTATTATCCTATGACCCTGCCACATCCCC-CTCTCTGAGAAACACCCAAGAATGATCAATAAATACTAA-GGGA
P311q221
P311q232
P312q141
P319p1311
P319q1313
P31p311
P33p25
P33q272
P34p16
P34q323
P36p211
P36p221
P3HERV-K(I)
P3HERV-K(II)
P3K101
P3K102
P3K103
P3K104
P3K105
P3K106
P3K107
P3K108S
P3K109
P3K110
P3K113
P3K115
P3Xq28
P3chimpX
P510p14
P511q221
P511q232
P512q141
P519p1311
P519q1313
P51p311
P53p25
P53q272
P54p16
P54q323
P56p211
P56p221
P5HERV-K(I)
P5HERV-K(II)
P5K101
P5K102
P5K103
P5K104
P5K105
P5K106
P5K107
P5K108S
P5K109
P5K110
P5K113
P5K115
P5Xq28
P5chimpX

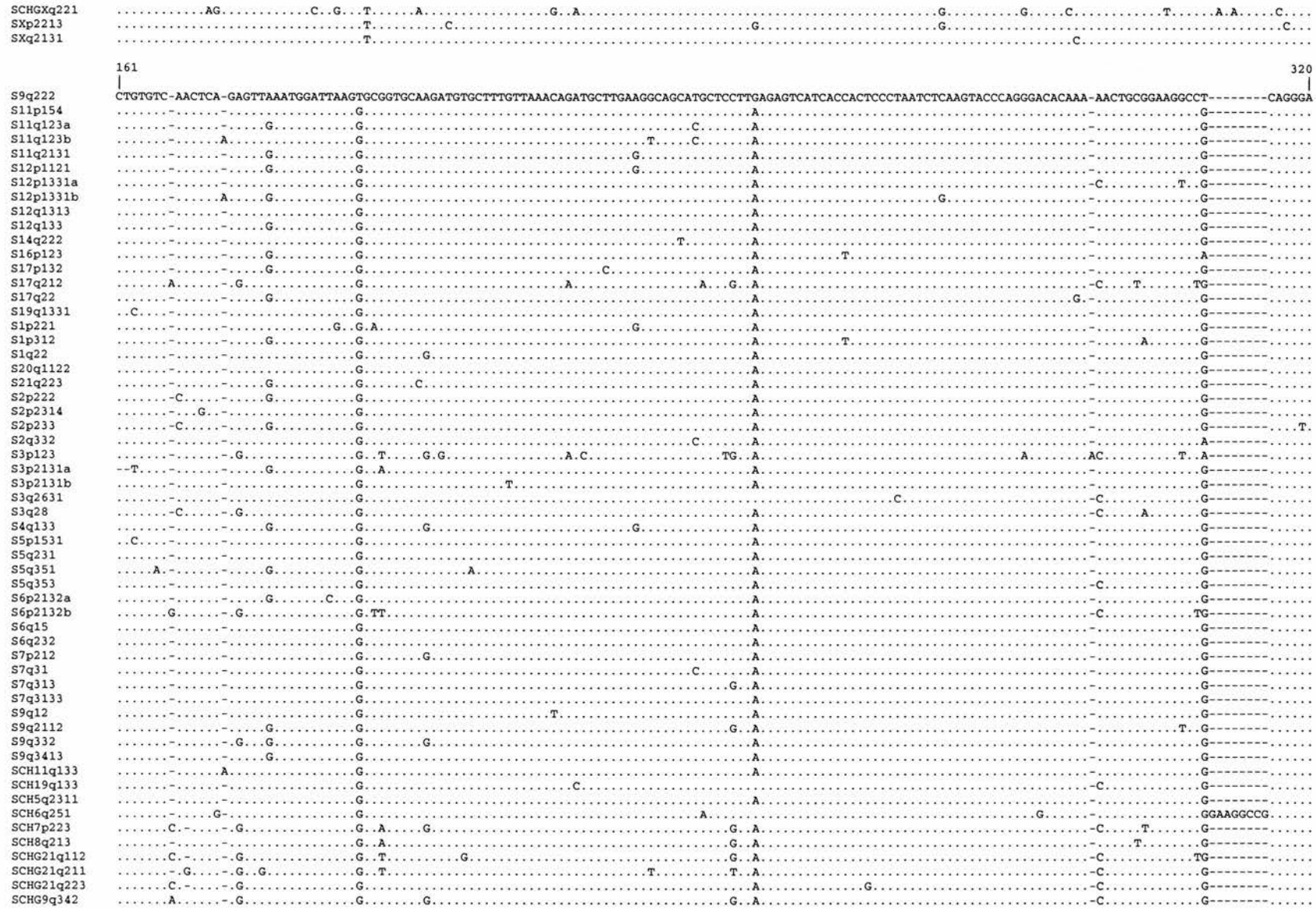

P3HERV-K(C19)
P521q211
961
1120
P310p14
P311q221
P311q232
P312q141
P319p1311
P319q1313
P31p311
P33p25
P33q272
P34p16
P34q323
P36p211
P36p221
P3HERV-K(I)
P3HERV-K(II)
P3K101
P3K102
P3K103
P3K104
P3K105
P3K106
P3K107
P3K108S
P3K109
P3K110
P3K113
P3K115
P3Xq28
P3chimpX
P510p14
P511q221
P511q232
P512q141
P519p1311
P519q1313
P51p311
P53p25
P53q272
P54p16
P54q323
P56p211
P56p221
P5HERV-K(I)
P5HERV-K(II)
P5K101
P5K102
P5K103
P5K104
P5K105
P5K106
P5K107
P5K108S
P5K109
P5K110

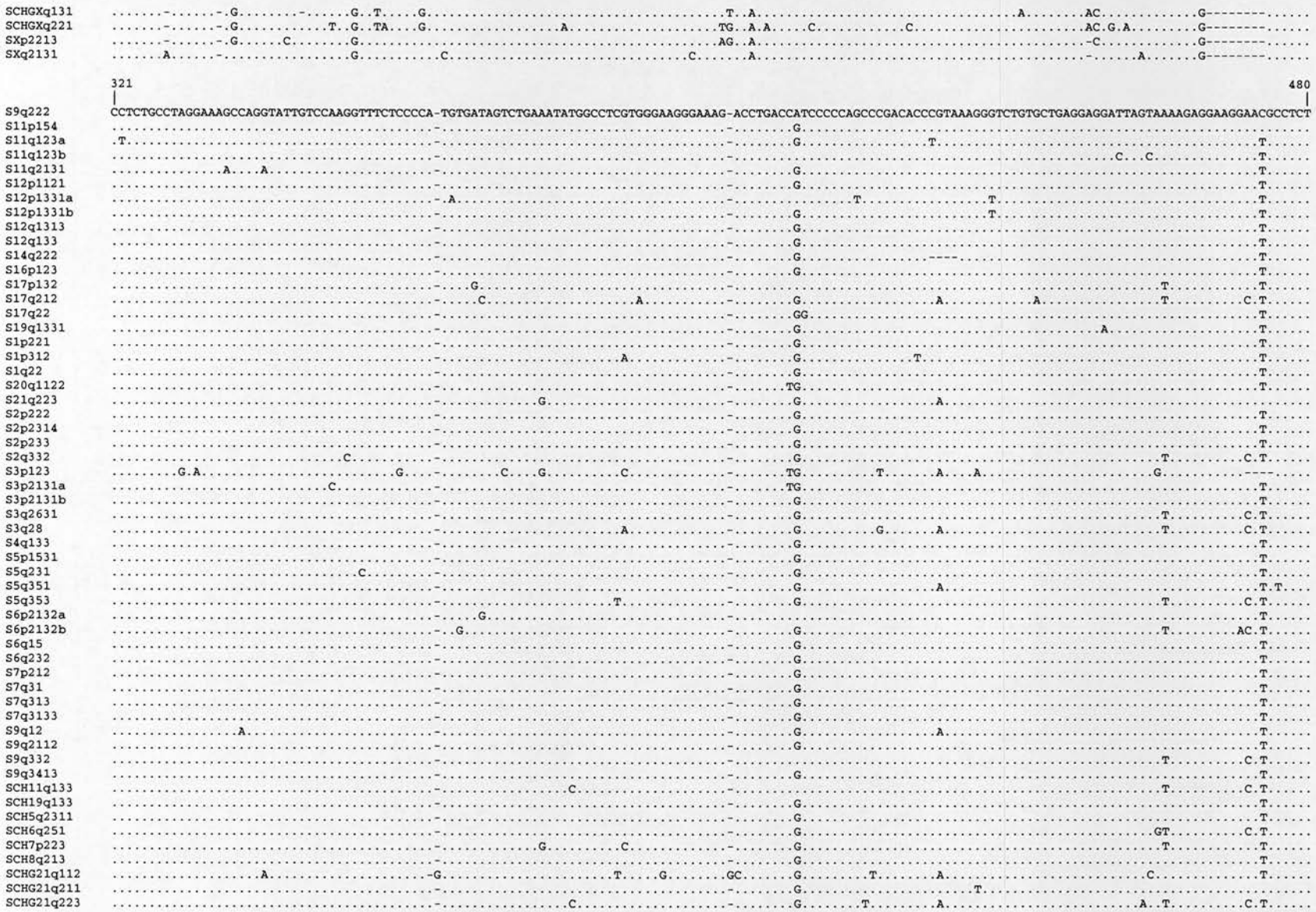

1121
1170
P5K113
P5K115
P5Xq28
P5chimpX
P3HERV-K(C19)
P521q211
P310p14
P311q221
P311q232
P312q141
P319p1311
P319q1313
P31p311
P33p25
P33q272
P34p16
P34q323
P36p211
P36p221
P3HERV-K(I)
P3HERV-K(II)
P3K101
P3K102
P3K103
P3K104
P3K105
P3K106
P3K107
P3K108S
P3K109
P3K110
P3K113
P3K115
P3Xq28
P3chimpX
P510p14
P511q221
P511q232
P512q141
P519p1311
P519q1313
P51p311
P53p25
P53q272
P54p16
P54q323
P56p211
P56p221
P5HERV-K(I)
P5HERV-K(II)
P5K101
P5K102
P5K103
P5K104
P5K105
P5K106
P5K107
P5K108S
P5K109

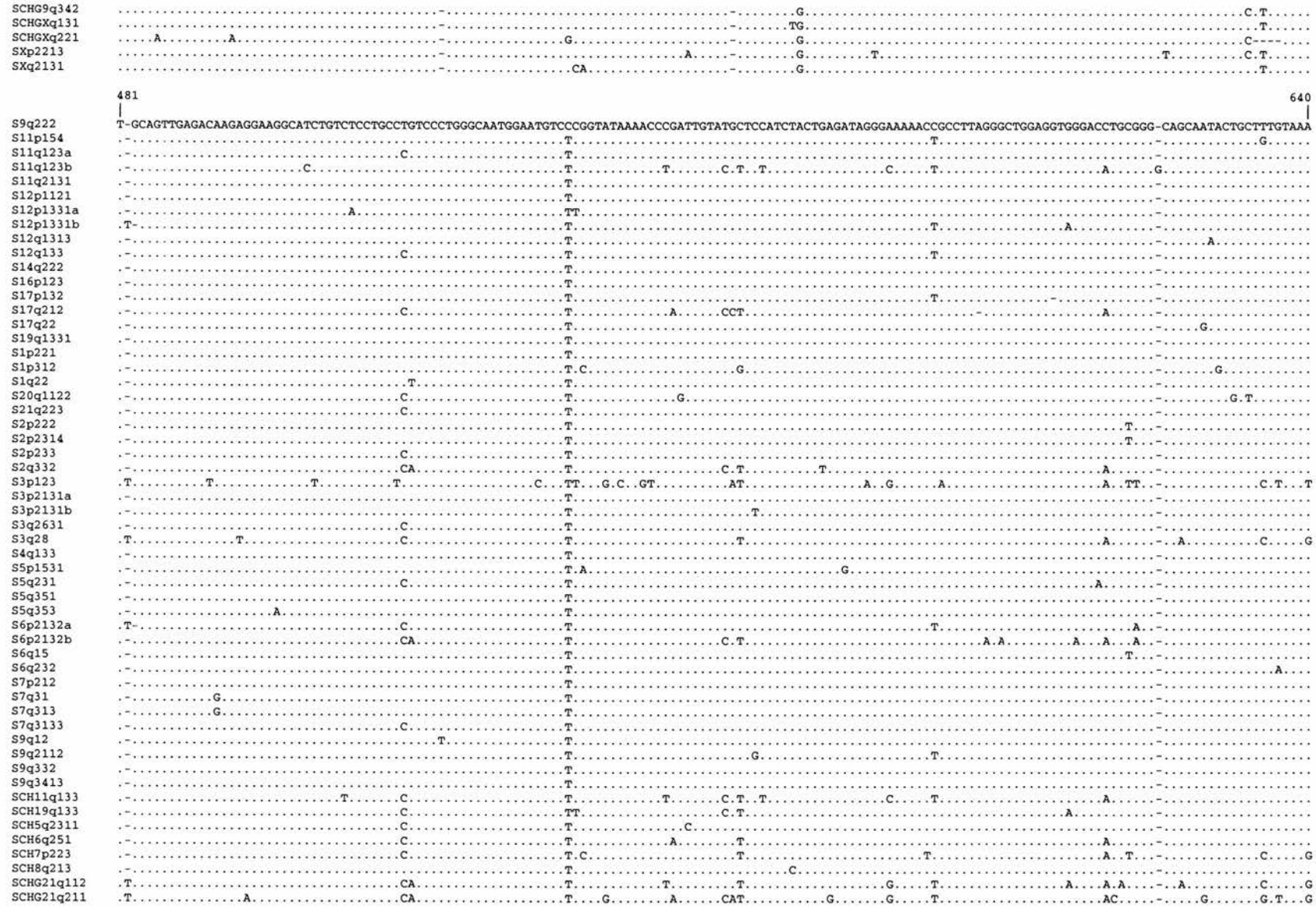
```
P5K110          .....--.....-........C.....G.....T.C..............
P5K113          .....--.....-....A...C..G.C......T.C..............
P5K115          .....--.....-....A...C..G.C......T.CA.............
P5Xq28          ..A..-.......-.TC.....C..G.C......G.C......T.......
P5chimpX        ..A..-.......-.TC.....C..G.C......G........T.......
P3HERV-K(C19)   .....--.....-....A...C..G.C......T.C..............
P521q211        .....--....C-.T......C..G........T.C..............
```

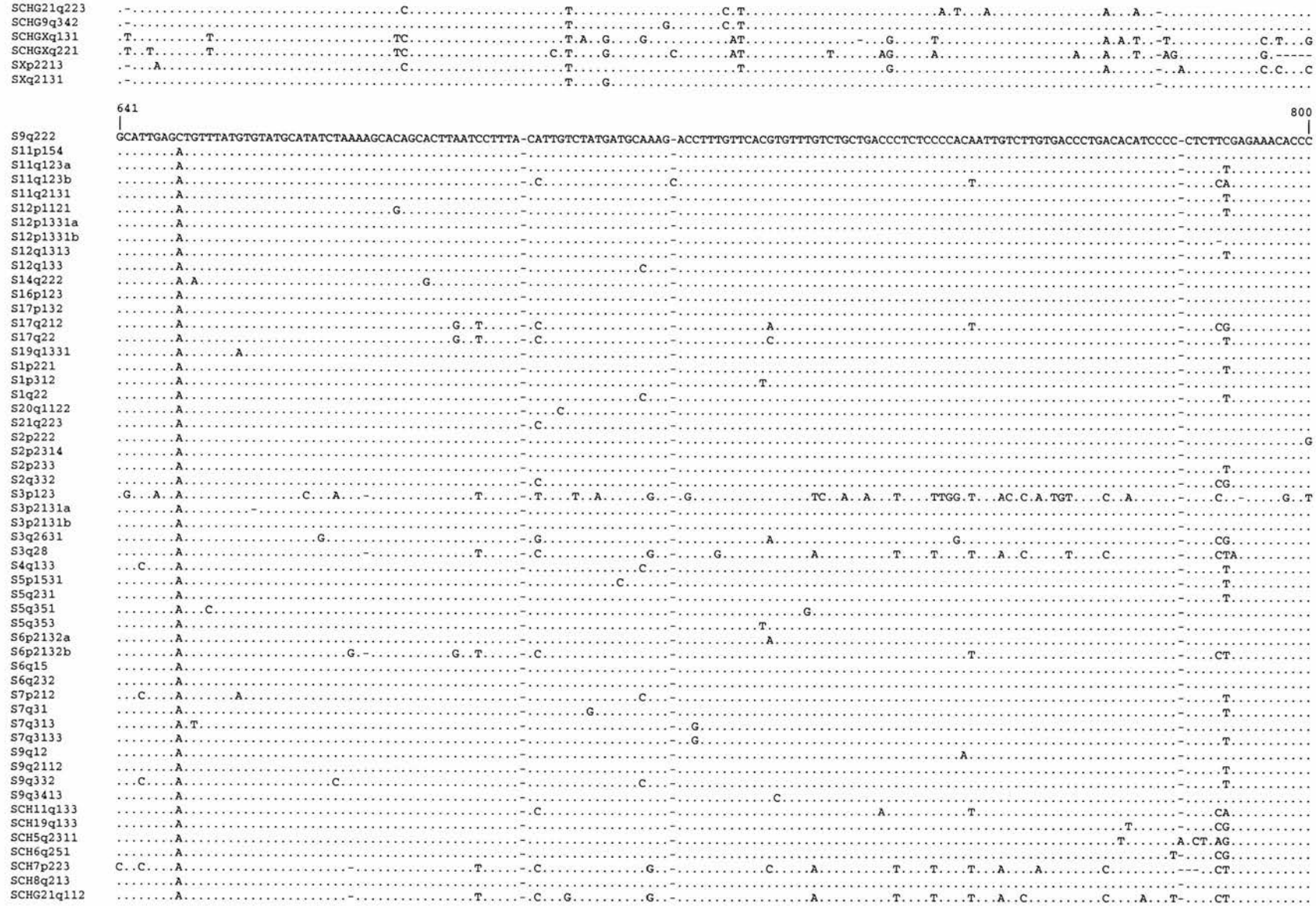
A.3 HERV-K(HML-2) Alignment of Solitary LTRs

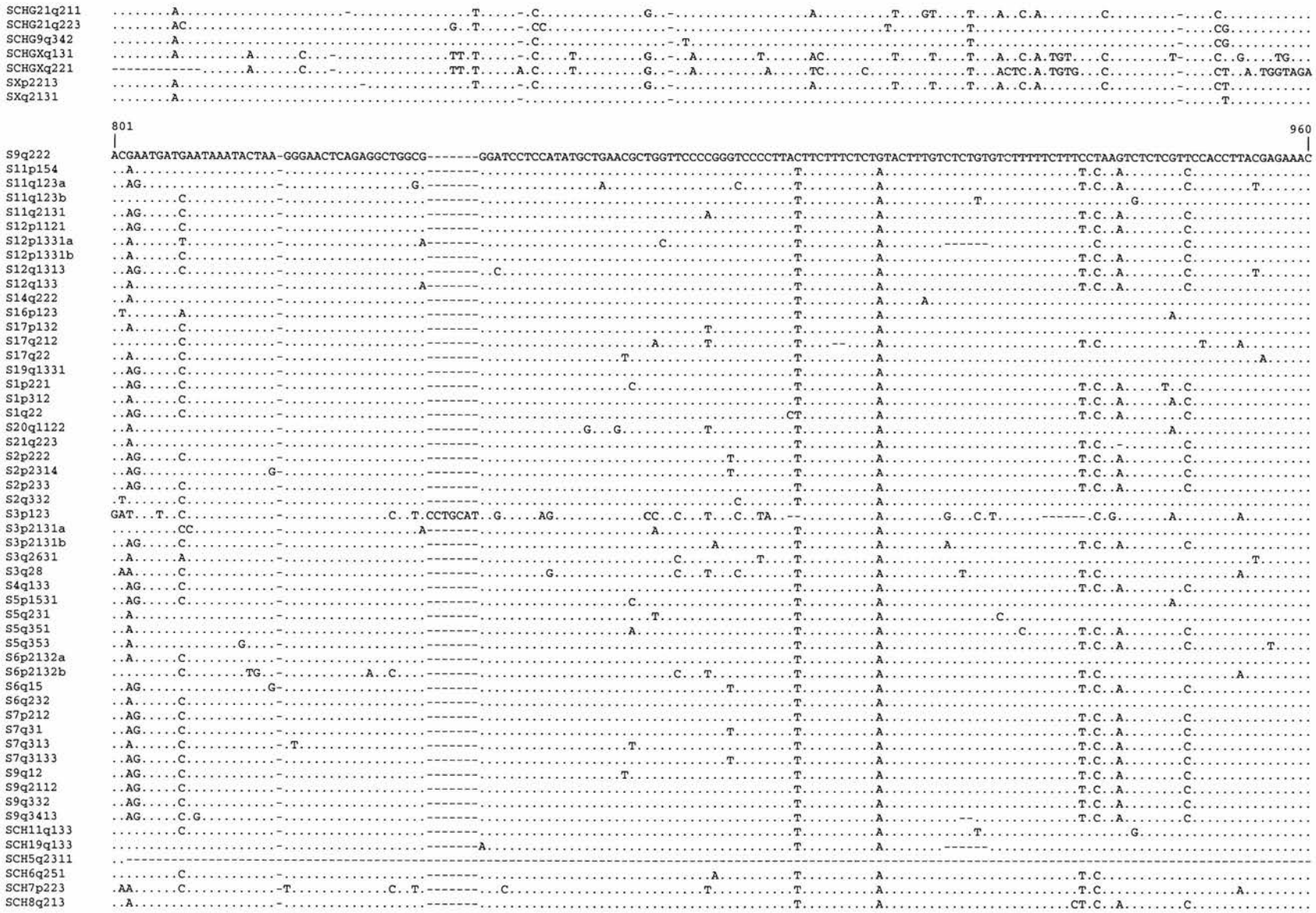


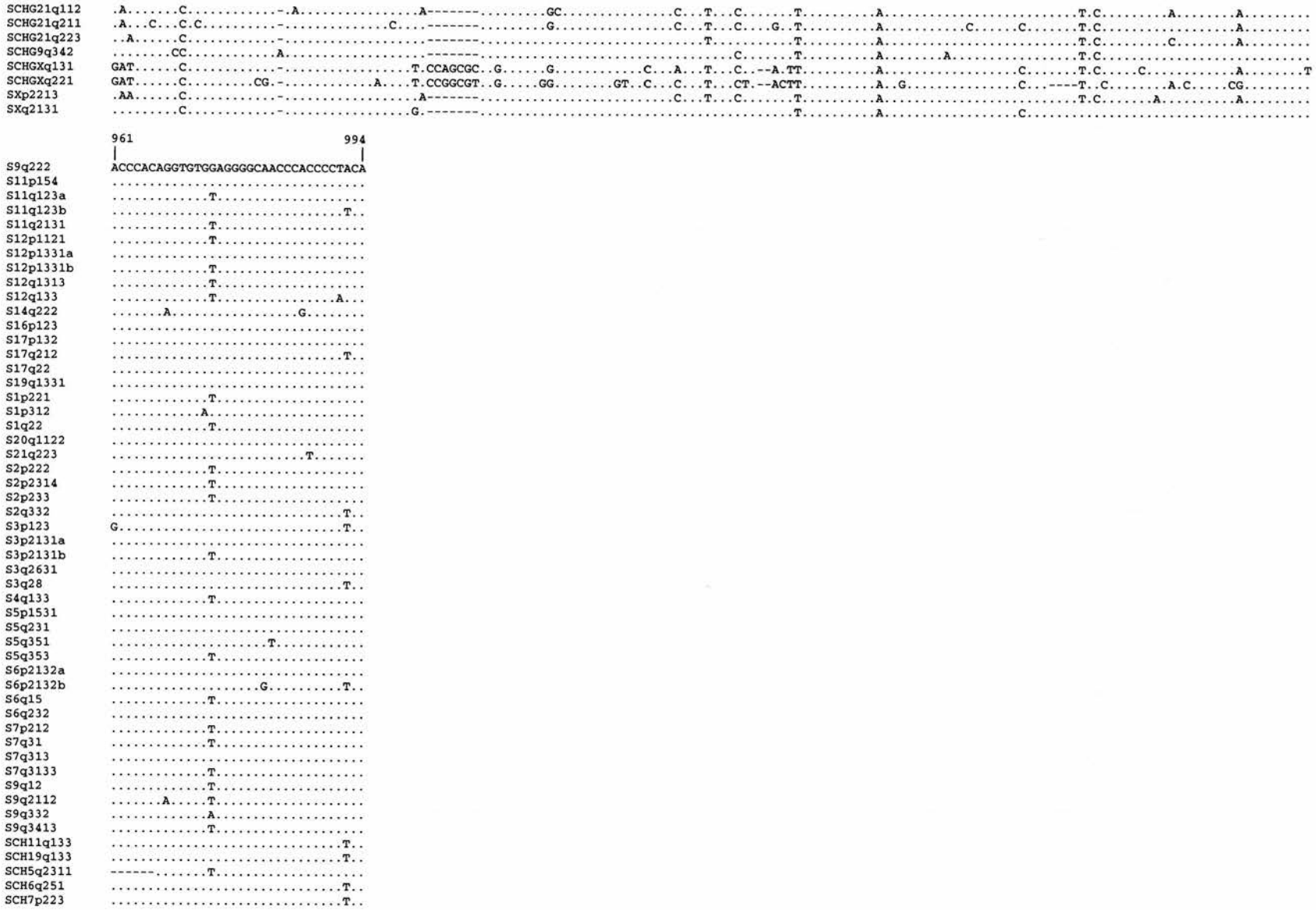

SCHGXq221
SXp2213
SXq2131
161
320
S9q222 CTGTGTC-AACTCA-GAGTTAAATGGATTAAGTGCGGTGCAAGATGTGCTTTGTTAAACAGATGCTTGAAGGCAGCATGCTCCTTGAGAGTCATCACCACTCCCTAATCTCAAGTACCCAGGGACACAAA-AACTGCGGAAGGCCT--------CAGGGA
S11p154
S11q123a
S11q123b
S11q2131
S12p1121
S12p1331a
S12p1331b
S12q1313
S12q133
S14q222
S16p123
S17p132
S17q212
S17q22
S19q1331
S1p221
S1p312
S1q22
S20q1122
S21q223
S2p222
S2p2314
S2p233
S2q332
S3p123
S3p2131a
S3p2131b
S3q2631
S3q28
S4q133
S5p1531
S5q231
S5q351
S5q353
S6p2132a
S6p2132b
S6q15
S6q232
S7p212
S7q31
S7q313
S7q3133
S9q12
S9q2112
S9q332
S9q3413
SCH11q133
SCH19q133
SCH5q2311
SCH6q251
SCH7p223
SCH8q213
SCHG21q112
SCHG21q211
SCHG21q223
SCHG9q342


SCHGXq131
SCHGXq221
SXp2213
SXq2131
321
480
S9q222
CCTCTGCCTAGGAAAGCCAGGTATTGTCCAAGGTTTCTCCCCA-TGTGATAGTCTGAAATATGGCCTCGTGGGAAGGGAAAG-ACCTGACCATCCCCCAGCCCGACACCCGTAAAGGGTCTGTGCTGAGGAGGATTAGTAAAAGAGGAAGGAACGCCTCT
S11p154
S11q123a
S11q123b
S11q2131
S12p1121
S12p1331a
S12p1331b
S12q1313
S12q133
S14q222
S16p123
S17p132
S17q212
S17q22
S19q1331
S1p221
S1p312
S1q22
S20q1122
S21q223
S2p222
S2p2314
S2p233
S2q332
S3p123
S3p2131a
S3p2131b
S3q2631
S3q28
S4q133
S5p1531
S5q231
S5q351
S5q353
S6p2132a
S6p2132b
S6q15
S6q232
S7p212
S7q31
S7q313
S7q3133
S9q12
S9q2112
S9q332
S9q3413
SCH11q133
SCH19q133
SCH5q2311
SCH6q251
SCH7p223
SCH8q213
SCHG21q112
SCHG21q211
SCHG21q223


481
640
S9q222 T-GCAGTTGAGACAAGAGGAAGGCATCTGTCTCCTGCCTGTCCCTGGGCAATGGAATGTCCCGGTATAAAACCCGATTGTATGCTCCATCTACTGAGATAGGGAAAAACCGCCTTAGGGCTGGAGGTGGGACCTGCGGG-CAGCAATACTGCTTTGTAAA
S11p154
S11q123a
S11q123b
S11q2131
S12p1121
S12p1331a
S12p1331b
S12q1313
S12q133
S14q222
S16p123
S17p132
S17q212
S17q22
S19q1331
S1p221
S1p312
S1q22
S20q1122
S21q223
S2p222
S2p2314
S2p233
S2q332
S3p123
S3p2131a
S3p2131b
S3q2631
S3q28
S4q133
S5p1531
S5q231
S5q351
S5q353
S6p2132a
S6p2132b
S6q15
S6q232
S7p212
S7q31
S7q313
S7q3133
S9q12
S9q2112
S9q332
S9q3413
SCH11q133
SCH19q133
SCH5q2311
SCH6q251
SCH7p223
SCH8q213
SCHG21q112
SCHG21q211
SCHG9q342
SCHGXq131
SCHGXq221
SXp2213
SXq2131

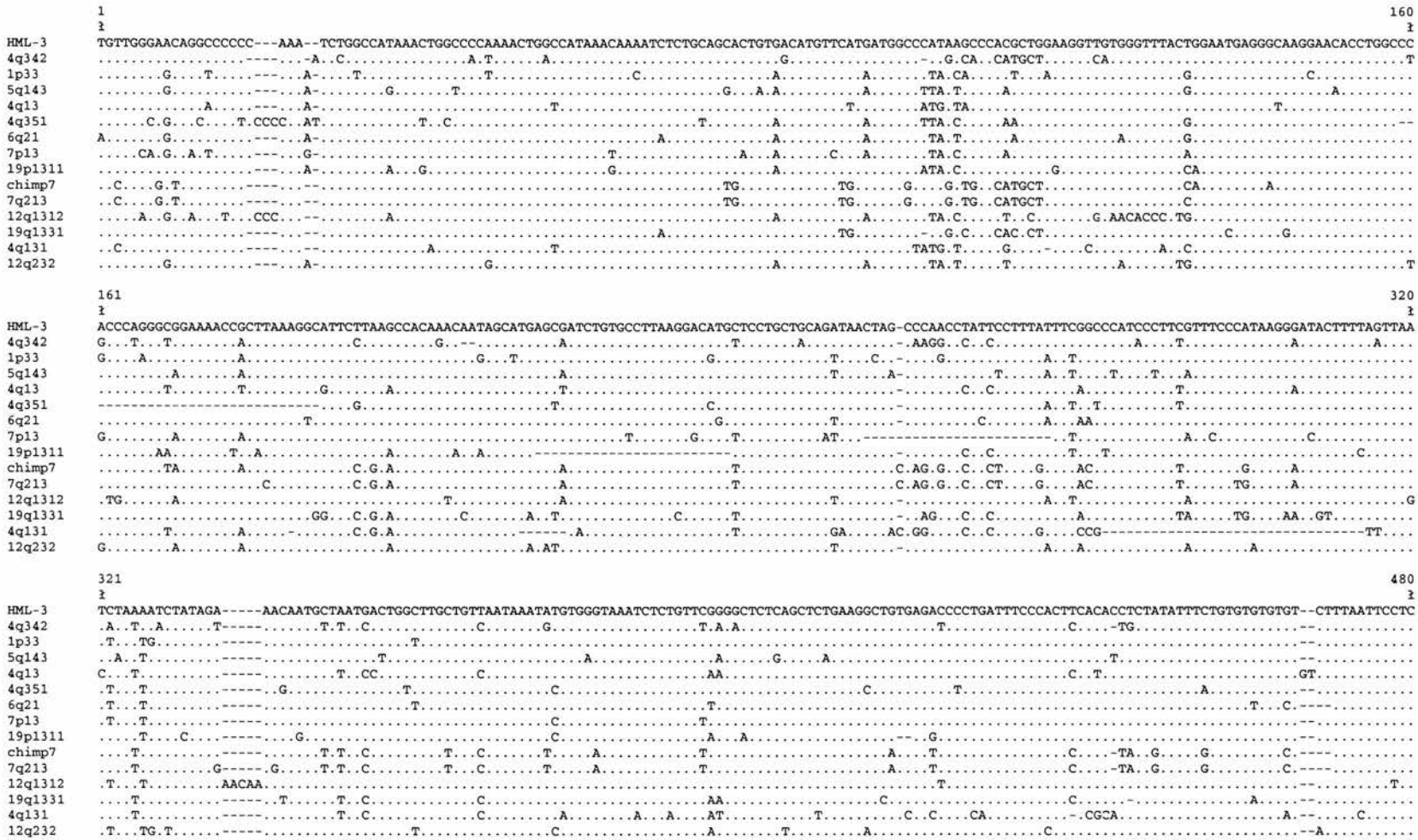

641
800
S9q222 GCATTGAGCTGTTTATGTGTATGCATATCTAAAAGCACAGCACTTAATCCTTTA-CATTGTCTATGATGCAAAG-ACCTTTGTTCACGTGTTTGTCTGCTGACCCTCTCCCCACAATTGTCTTGTGACCCTGACACATCCCC-CTCTTCGAGAAACACCC
S11p154
S11q123a
S11q123b
S11q2131
S12p1121
S12p1331a
S12p1331b
S12q1313
S12q133
S14q222
S16p123
S17p132
S17q212
S17q22
S19q1331
S1p221
S1p312
S1q22
S20q1122
S21q223
S2p222
S2p2314
S2p233
S2q332
S3p123
S3p2131a
S3p2131b
S3q2631
S3q28
S4q133
S5p1531
S5q231
S5q351
S5q353
S6p2132a
S6p2132b
S6q15
S6q232
S7p212
S7q31
S7q313
S7q3133
S9q12
S9q2112
S9q332
S9q3413
SCH11q133
SCH19q133
SCH5q2311
SCH6q251
SCH7p223
SCH8q213
SCHG21q112
SCHG21q223
SCHG9q342
SCHGXq131
SCHGXq221
SXp2213
SXq2131

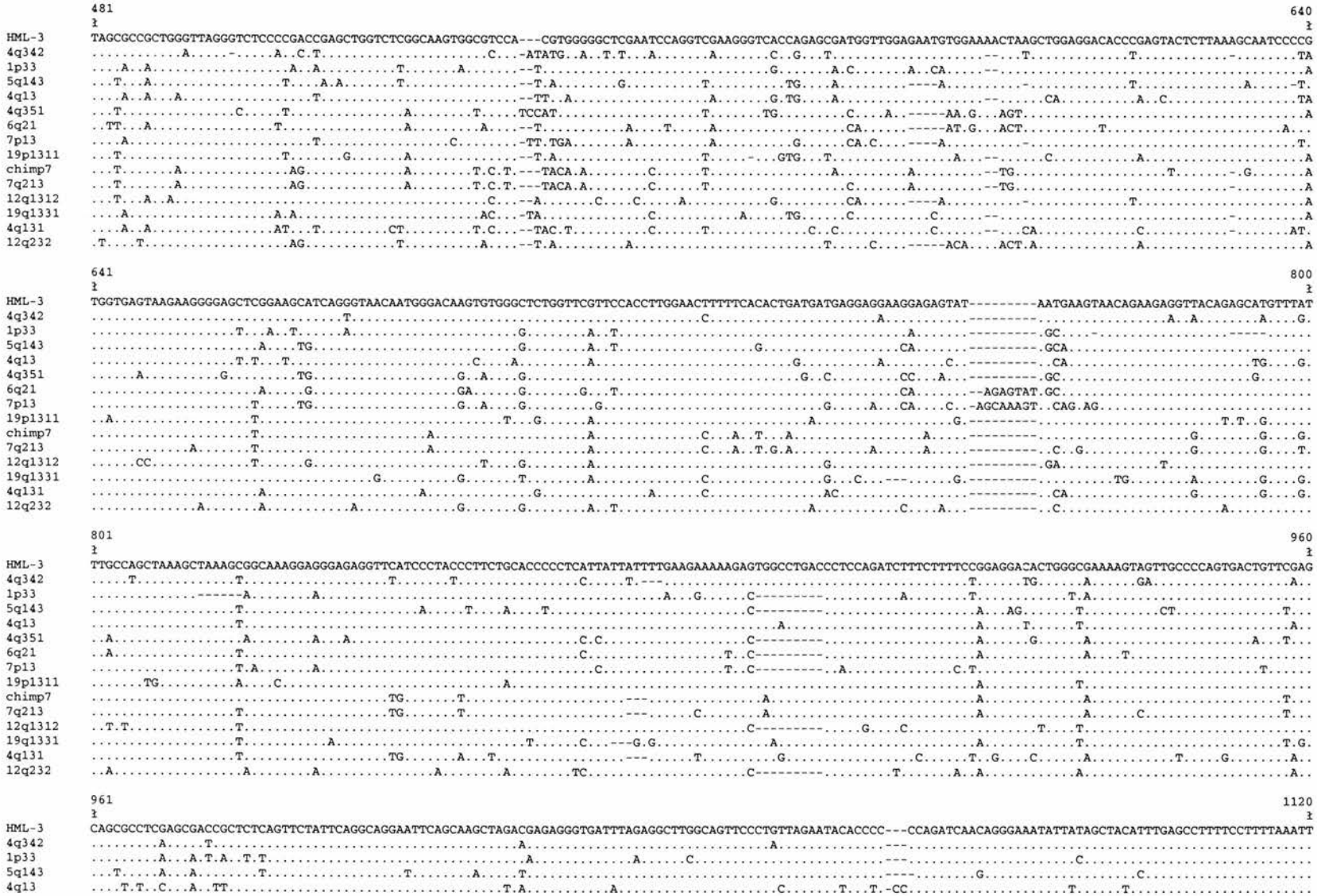

SCHG21q211
SCHG21q223
SCHG9q342
SCHGXq131
SCHGXq221
SXp2213
SXq2131
801
960
S9q222
ACGAATGATGAATAAATACTAA-GGGAACTCAGAGGCTGGCG-------GGATCCTCCATATGCTGAACGCTGGTTCCCCGGGTCCCCTTACTTCTTTCTCTGTACTTTGTCTCTGTGTCTTTTTCTTTCCTAAGTCTCTCGTTCCACCTTACGAGAAAC
S11p154
S11q123a
S11q123b
S11q2131
S12p1121
S12p1331a
S12p1331b
S12q1313
S12q133
S14q222
S16p123
S17p132
S17q212
S17q22
S19q1331
S1p221
S1p312
S1q22
S20q1122
S21q223
S2p222
S2p2314
S2p233
S2q332
S3p123
S3p2131a
S3p2131b
S3q2631
S3q28
S4q133
S5p1531
S5q231
S5q351
S5q353
S6p2132a
S6p2132b
S6q15
S6q232
S7p212
S7q31
S7q313
S7q3133
S9q12
S9q2112
S9q332
S9q3413
SCH11q133
SCH19q133
SCH5q2311
SCH6q251
SCH7p223
SCH8q213

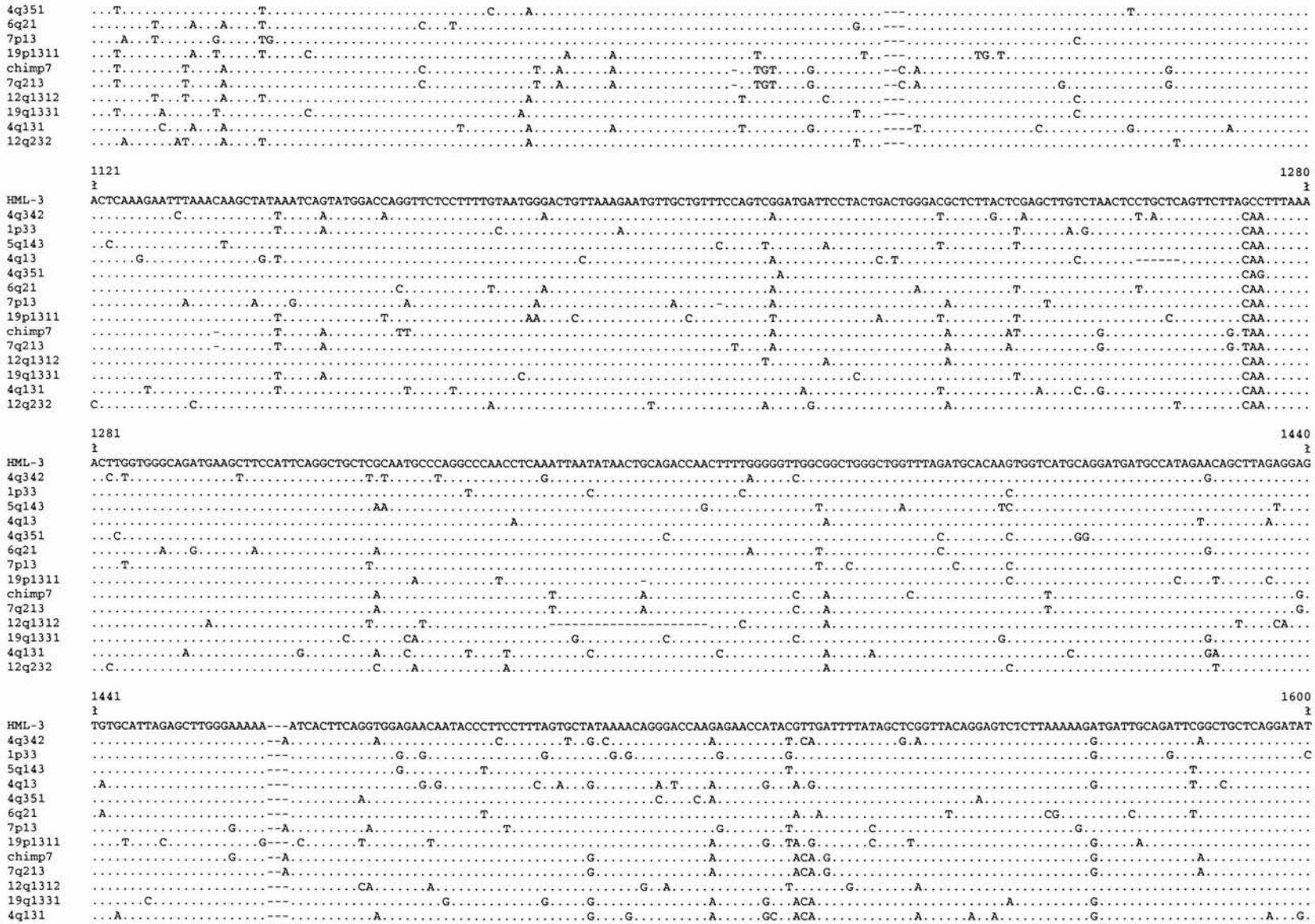

961
994
S9q222 ACCCACAGGTGTGGAGGGGCAACCCACCCCTACA
S11p154
S11q123a
S11q123b
S11q2131
S12p1121
S12p1331a
S12p1331b
S12q1313
S12q133
S14q222
S16p123
S17p132
S17q212
S17q22
S19q1331
S1p221
S1p312
S1q22
S20q1122
S21q223
S2p222
S2p2314
S2p233
S2q332
S3p123
S3p2131a
S3p2131b
S3q2631
S3q28
S4q133
S5p1531
S5q231
S5q351
S5q353
S6p2132a
S6p2132b
S6q15
S6q232
S7p212
S7q31
S7q313
S7q3133
S9q12
S9q2112
S9q332
S9q3413
SCH11q133
SCH19q133
SCH5q2311
SCH6q251
SCH7p223
SCHG21q112
SCHG21q211
SCHG21q223
SCHG9q342
SCHGXq131
SCHGXq221
SXp2213
SXq2131

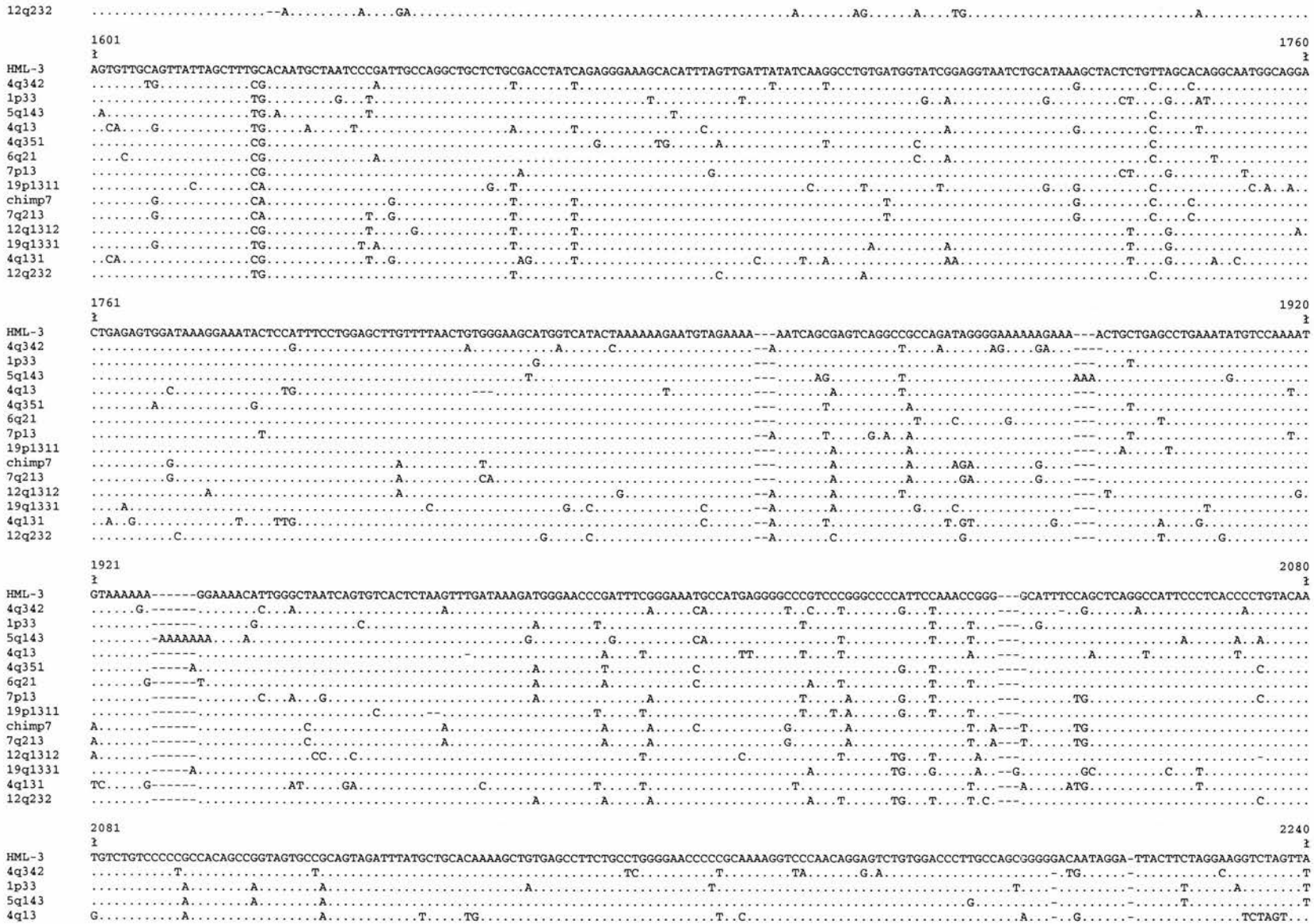
```
SCH8q213        ............T.....................
SCHG21q112      ...............................T..
SCHG21q211      ......................A....A...T.G
SCHG21q223      ...............................T..
SCHG9q342       ...............................T..
SCHGXq131       G.........................T....T..
SCHGXq221       ......................GG........T..
SXp2213         ...............................T..
SXq2131         ......................A...........
```

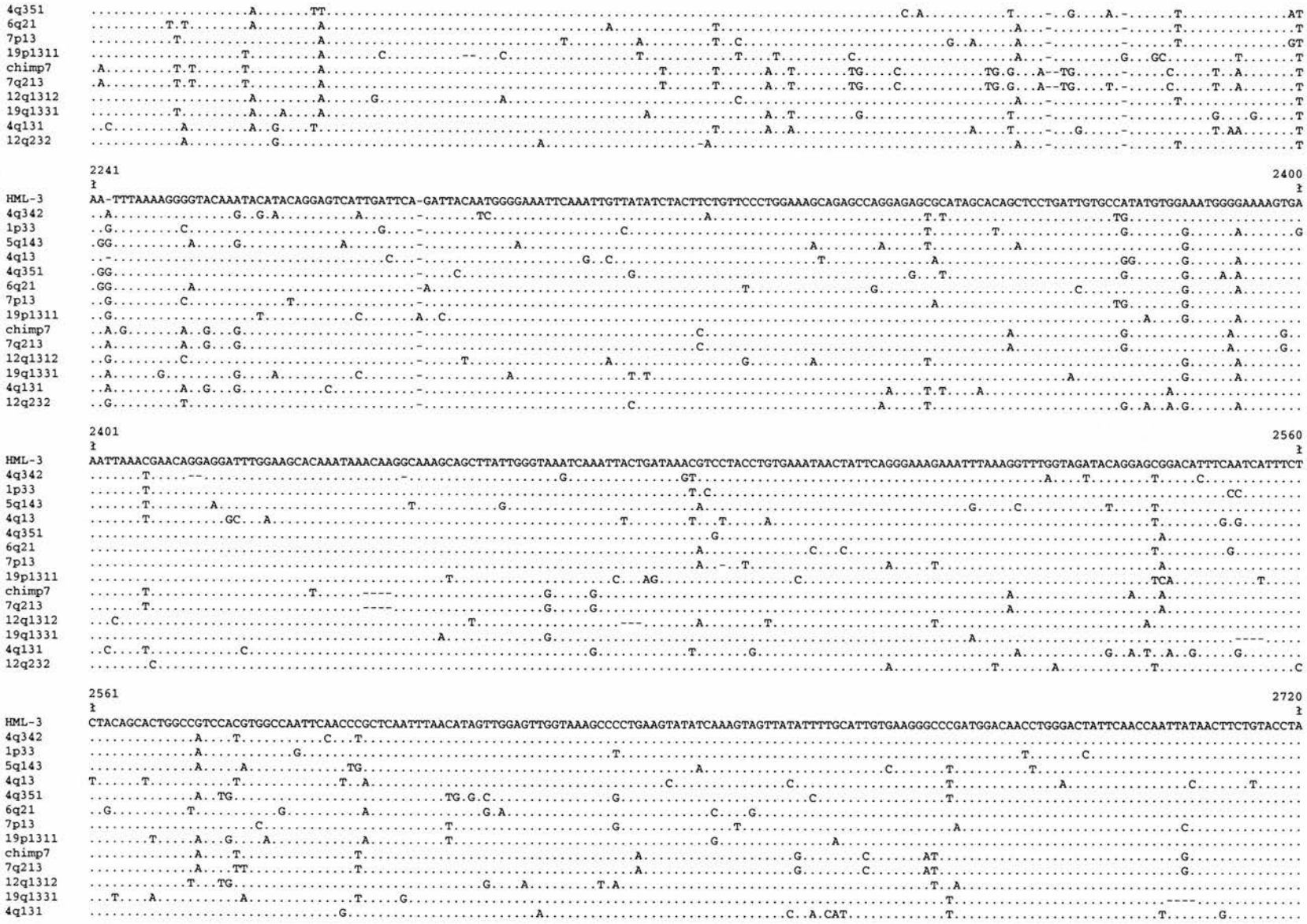
A.4 HERV-K(HML-3) Alignment of complete proviruses used in study

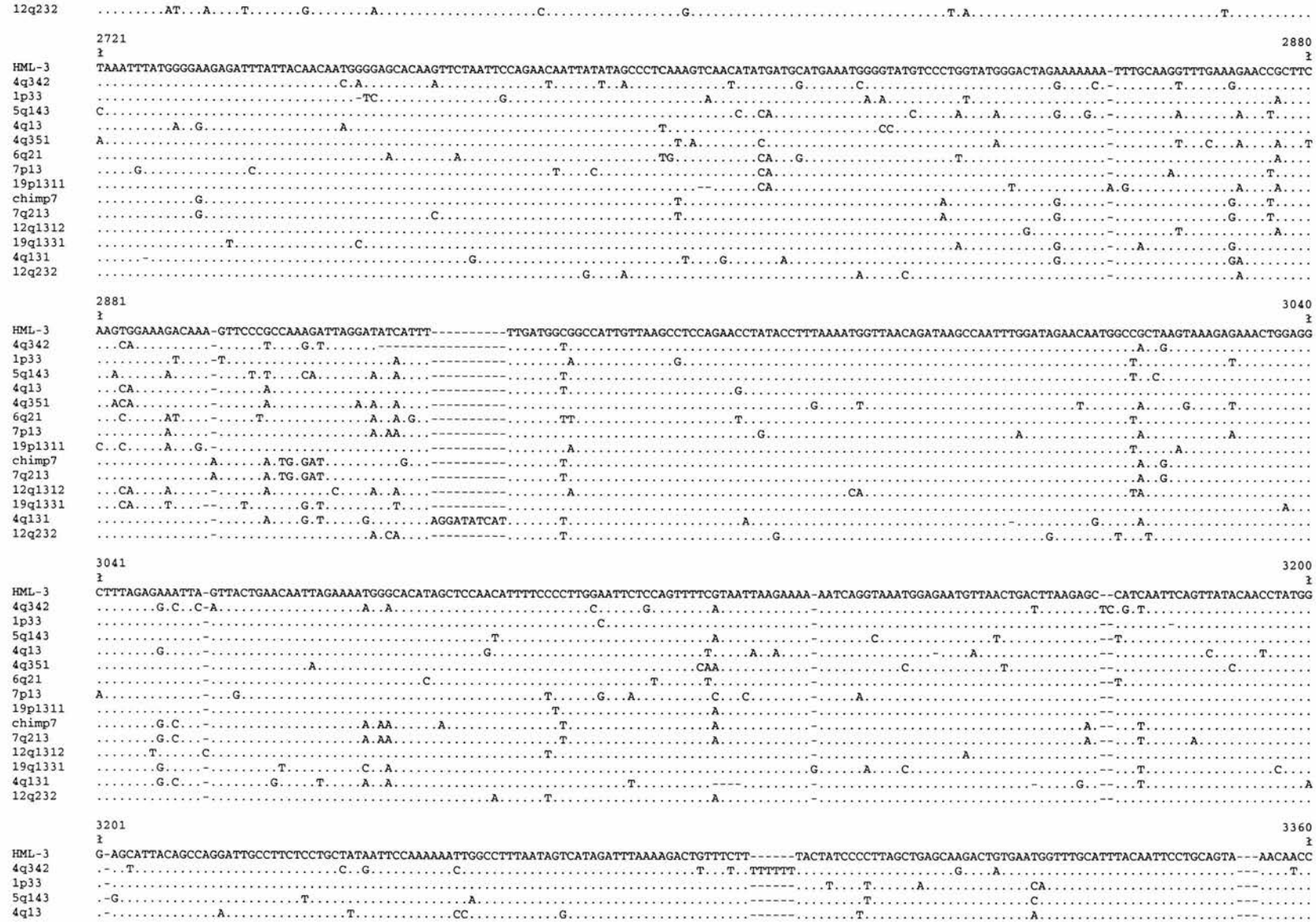


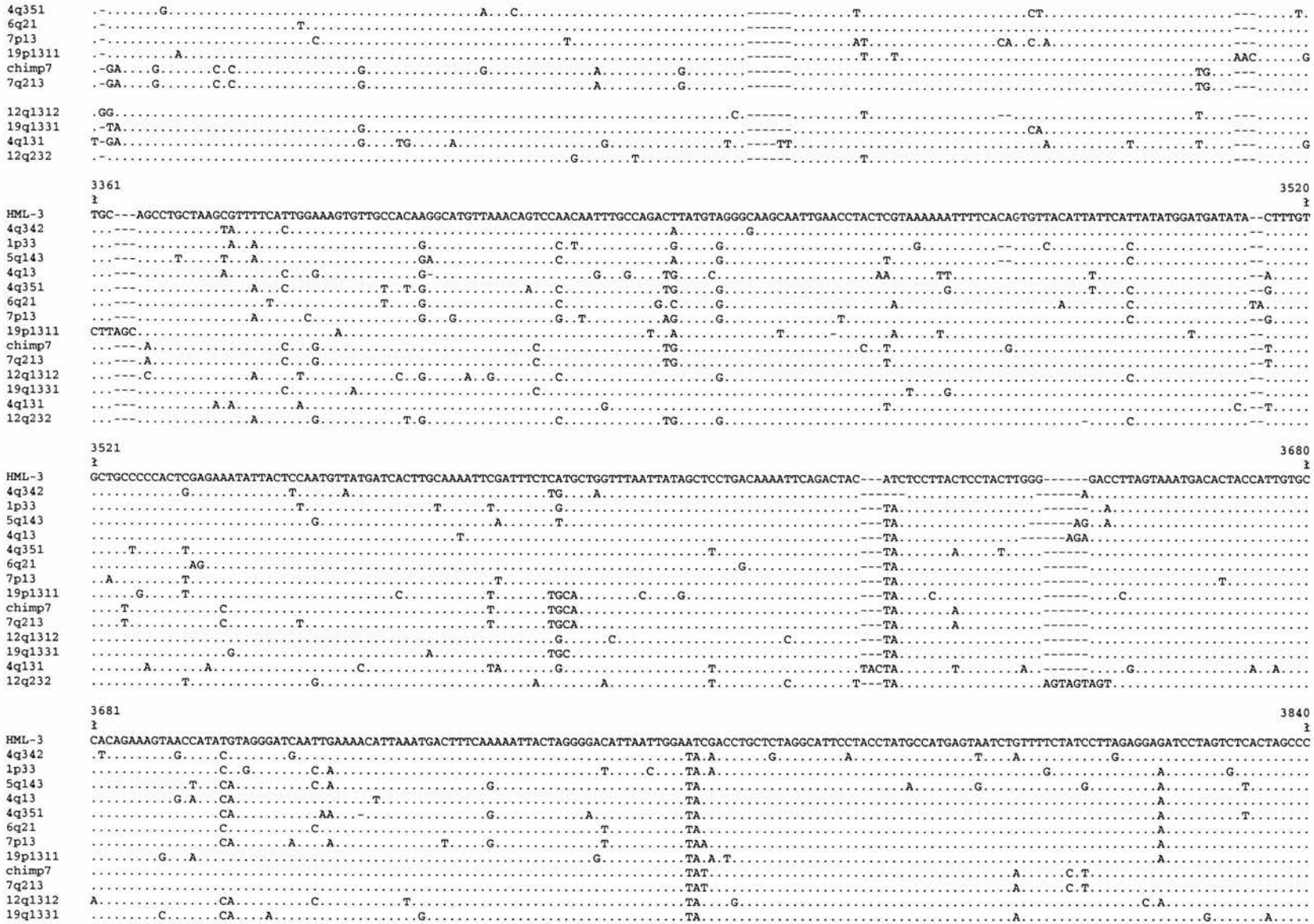

481
640
HML-3 TAGCGCCGCTGGGTTAGGGTCTCCCCGACCGAGCTGGTCTCGGCAAGTGGCGTCCA---CGTGGGGGCTCGAATCCAGGTCGAAGGGTCACCAGAGCGATGGTTGGAGAATGTGGAAAACTAAGCTGGAGGACACCCGAGTACTCTTAAAGCAATCCCCG
4q342
1p33
5q143
4q13
4q351
6q21
7p13
19p1311
chimp7
7q213
12q1312
19q1331
4q131
12q232
641
800
HML-3 TGGTGAGTAAGAAGGGGAGCTCGGAAGCATCAGGGTAACAATGGGACAAGTGTGGGCTCTGGTTCGTTCCACCTTGGAACTTTTTCACACTGATGATGAGGAGGAAGGAGAGTAT---------AATGAAGTAACAGAAGAGGTTACAGAGCATGTTTAT
4q342
1p33
5q143
4q13
4q351
6q21
7p13
19p1311
chimp7
7q213
12q1312
19q1331
4q131
12q232
801
960
HML-3 TTGCCAGCTAAAGCTAAAGCGGCAAAGGAGGGAGAGGTTCATCCCTACCCTTCTGCACCCCCTCATTATTATTTTGAAGAAAAAGAGTGGCCTGACCCTCCAGATCTTTCTTTTCCGGAGGACACTGGGCGAAAAGTAGTTGCCCCAGTGACTGTTCGAG
4q342
1p33
5q143
4q13
4q351
6q21
7p13
19p1311
chimp7
7q213
12q1312
19q1331
4q131
12q232
961
1120
HML-3 CAGCGCCTCGAGCGACCGCTCTCAGTTCTATTCAGGCAGGAATTCAGCAAGCTAGACGAGAGGGTGATTTAGAGGCTTGGCAGTTCCCTGTTAGAATACACCCC---CCAGATCAACAGGGAAATATTATAGCTACATTTGAGCCTTTTCCTTTTAAATT
4q342
1p33
5q143
4q13

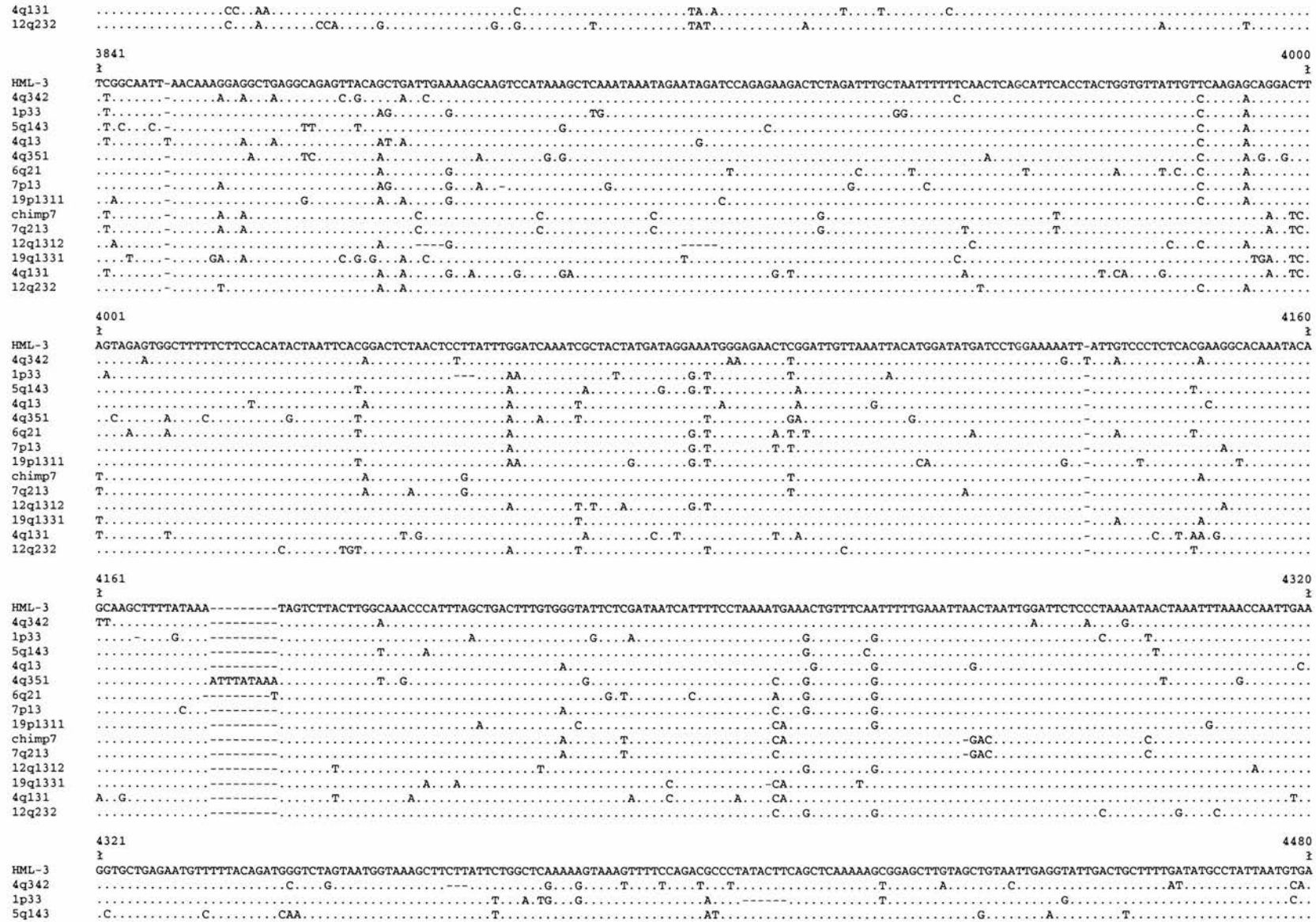

4q351
6q21
7p13
19p1311
chimp7
7q213
12q1312
19q1331
4q131
12q232
1121
1280
HML-3
ACTCAAAGAATTTAAACAAGCTATAAATCAGTATGGACCAGGTTCTCCTTTTGTAATGGGACTGTTAAAGAATGTTGCTGTTTCCAGTCGGATGATTCCTACTGACTGGGACGCTCTTACTCGAGCTTGTCTAACTCCTGCTCAGTTCTTAGCCTTTAAA
4q342
1p33
5q143
4q13
4q351
6q21
7p13
19p1311
chimp7
7q213
12q1312
19q1331
4q131
12q232
1281
1440
HML-3
ACTTGGTGGGCAGATGAAGCTTCCATTCAGGCTGCTCGCAATGCCCAGGCCCAACCTCAAATTAATATAACTGCAGACCAACTTTTGGGGGTTGGCGGCTGGGCTGGTTTAGATGCACAAGTGGTCATGCAGGATGATGCCATAGAACAGCTTAGAGGAG
4q342
1p33
5q143
4q13
4q351
6q21
7p13
19p1311
chimp7
7q213
12q1312
19q1331
4q131
12q232
1441
1600
HML-3
TGTGCATTAGAGCTTGGGAAAAA---ATCACTTCAGGTGGAGAACAATACCCTTCCTTTAGTGCTATAAAACAGGGACCAAGAGAACCATACGTTGATTTTATAGCTCGGTTACAGGAGTCTCTTAAAAAGATGATTGCAGATTCGGCTGCTCAGGATAT
4q342
1p33
5q143
4q13
4q351
6q21
7p13
19p1311
chimp7
7q213
12q1312
19q1331
4q131

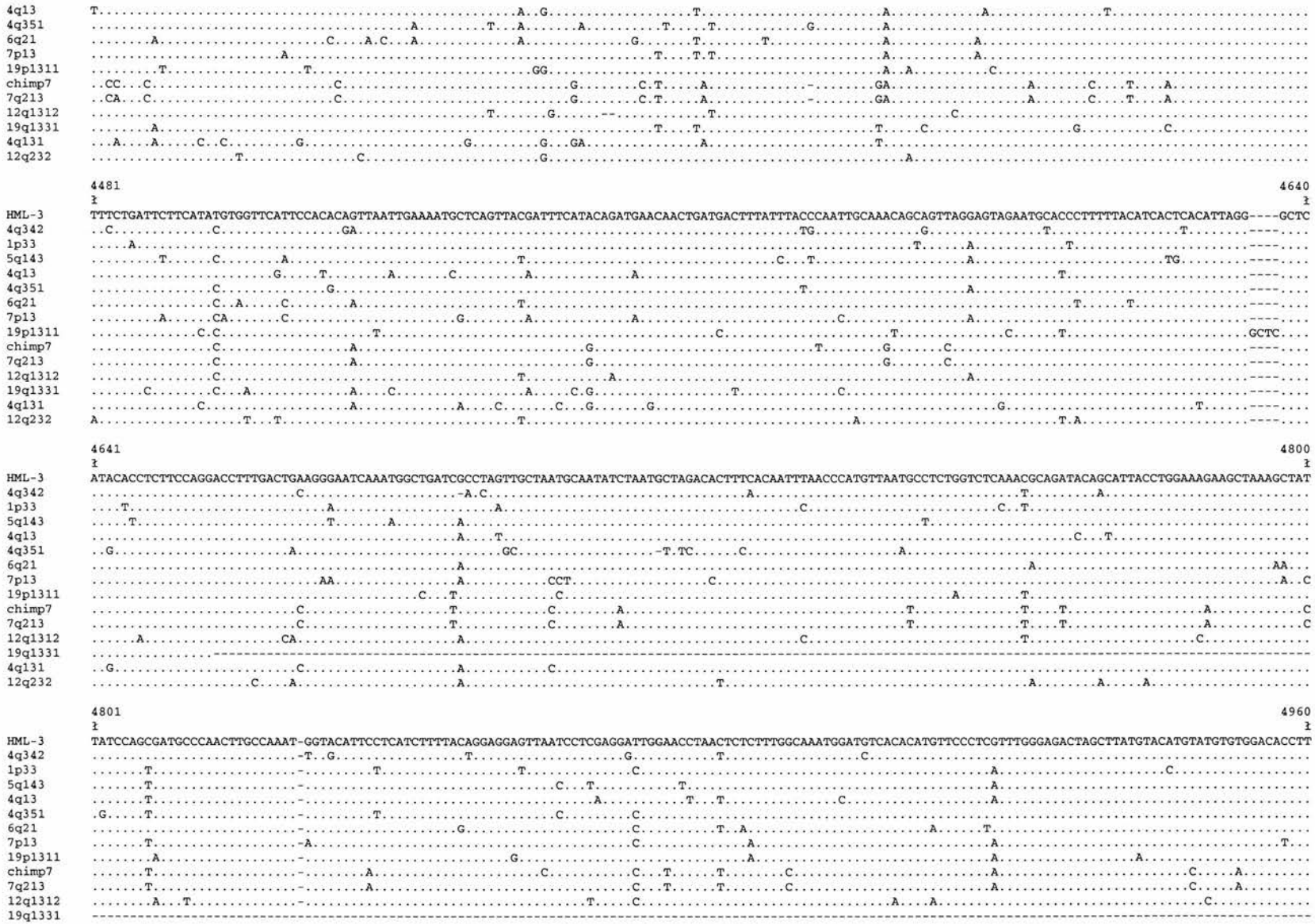

12q232
1601
1760
HML-3
4q342
1p33
5q143
4q13
4q351
6q21
7p13
19p1311
chimp7
7q213
12q1312
19q1331
4q131
12q232
1761
1920
1921
2080
2081
2240

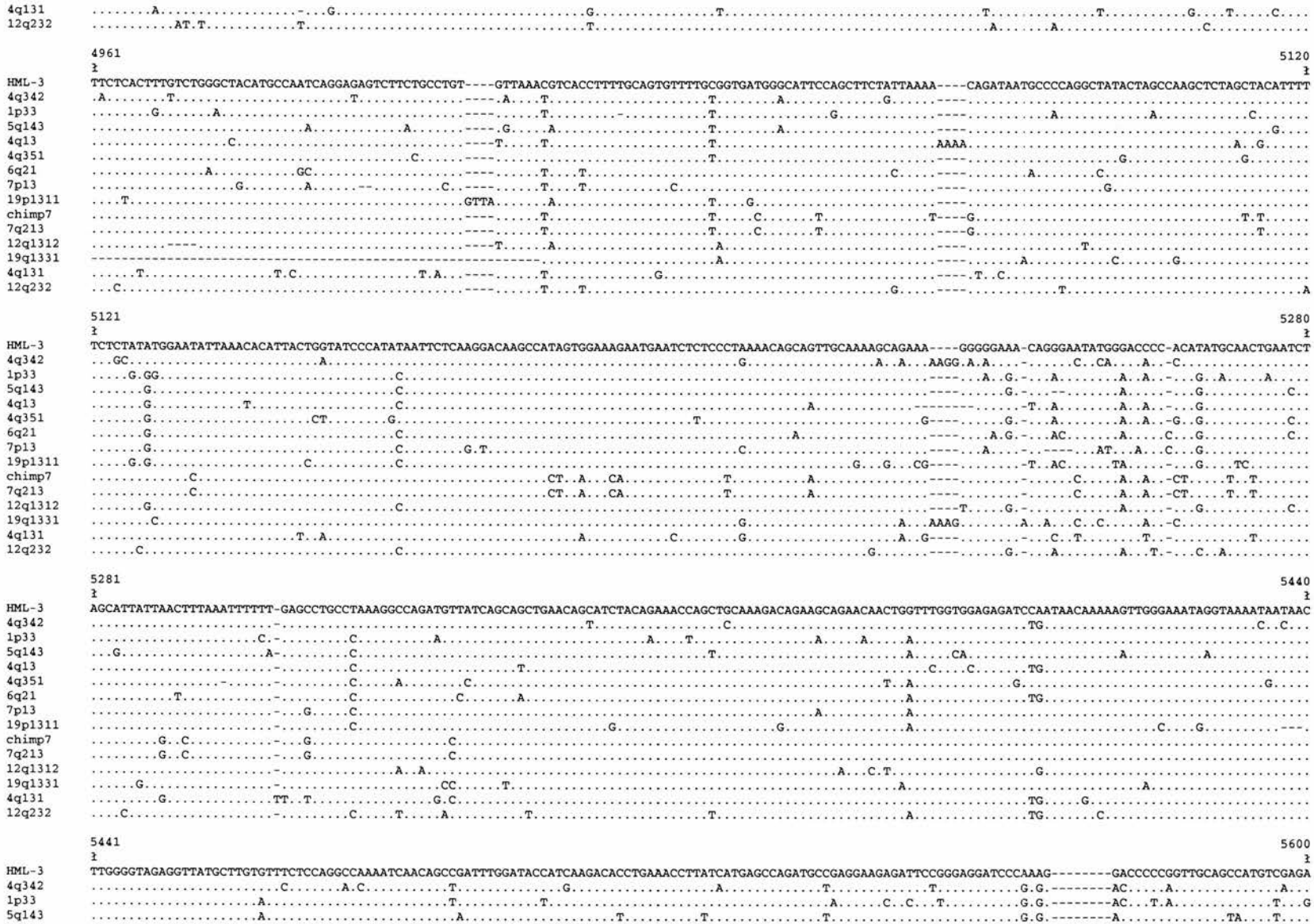

4q351
6q21
7p13
19p1311
chimp7
7q213
12q1312
19q1331
4q131
12q232
2241
2400
HML-3
AA-TTTAAAAGGGGTACAAATACATACAGGAGTCATTGATTCA-GATTACAATGGGGAAATTCAAATTGTTATATCTACTTCTGTTCCCTGGAAAGCAGAGCCAGGAGAGCGCATAGCACAGCTCCTGATTGTGCCATATGTGGAAATGGGGAAAAGTGA
4q342
1p33
5q143
4q13
4q351
6q21
7p13
19p1311
chimp7
7q213
12q1312
19q1331
4q131
12q232
2401
2560
HML-3
AATTAAACGAACAGGAGGATTTGGAAGCACAAATAAACAAGGCAAAGCAGCTTATTGGGTAAATCAAATTACTGATAAACGTCCTACCTGTGAAATAACTATTCAGGGAAAGAAATTTAAAGGTTTGGTAGATACAGGAGCGGACATTTCAATCATTTCT
4q342
1p33
5q143
4q13
4q351
6q21
7p13
19p1311
chimp7
7q213
12q1312
19q1331
4q131
12q232
2561
2720
HML-3
CTACAGCACTGGCCGTCCACGTGGCCAATTCAACCCGCTCAATTTAACATAGTTGGAGTTGGTAAAGCCCCTGAAGTATATCAAAGTAGTTATATTTTGCATTGTGAAGGGCCCGATGGACAACCTGGGACTATTCAACCAATTATAACTTCTGTACCTA
4q342
1p33
5q143
4q13
4q351
6q21
7p13
19p1311
chimp7
7q213
12q1312
19q1331
4q131

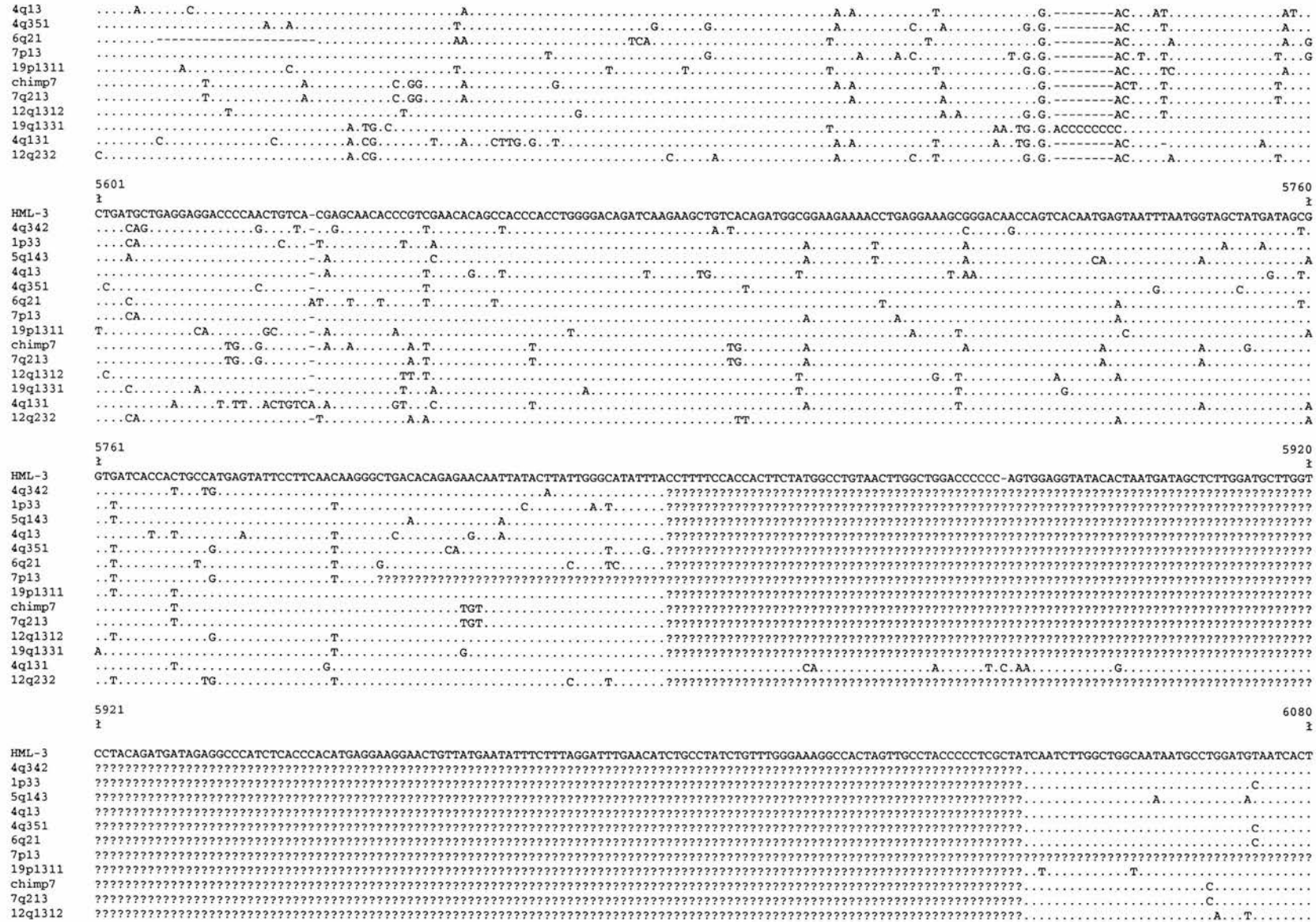

12q232
2721
2880
HML-3
4q342
1p33
5q143
4q13
4q351
6q21
7p13
19p1311
chimp7
7q213
12q1312
19q1331
4q131
12q232
2881
3040
HML-3
4q342
1p33
5q143
4q13
4q351
6q21
7p13
19p1311
chimp7
7q213
12q1312
19q1331
4q131
12q232
3041
3200
HML-3
4q342
1p33
5q143
4q13
4q351
6q21
7p13
19p1311
chimp7
7q213
12q1312
19q1331
4q131
12q232
3201
3360
HML-3
4q342
1p33
5q143
4q13

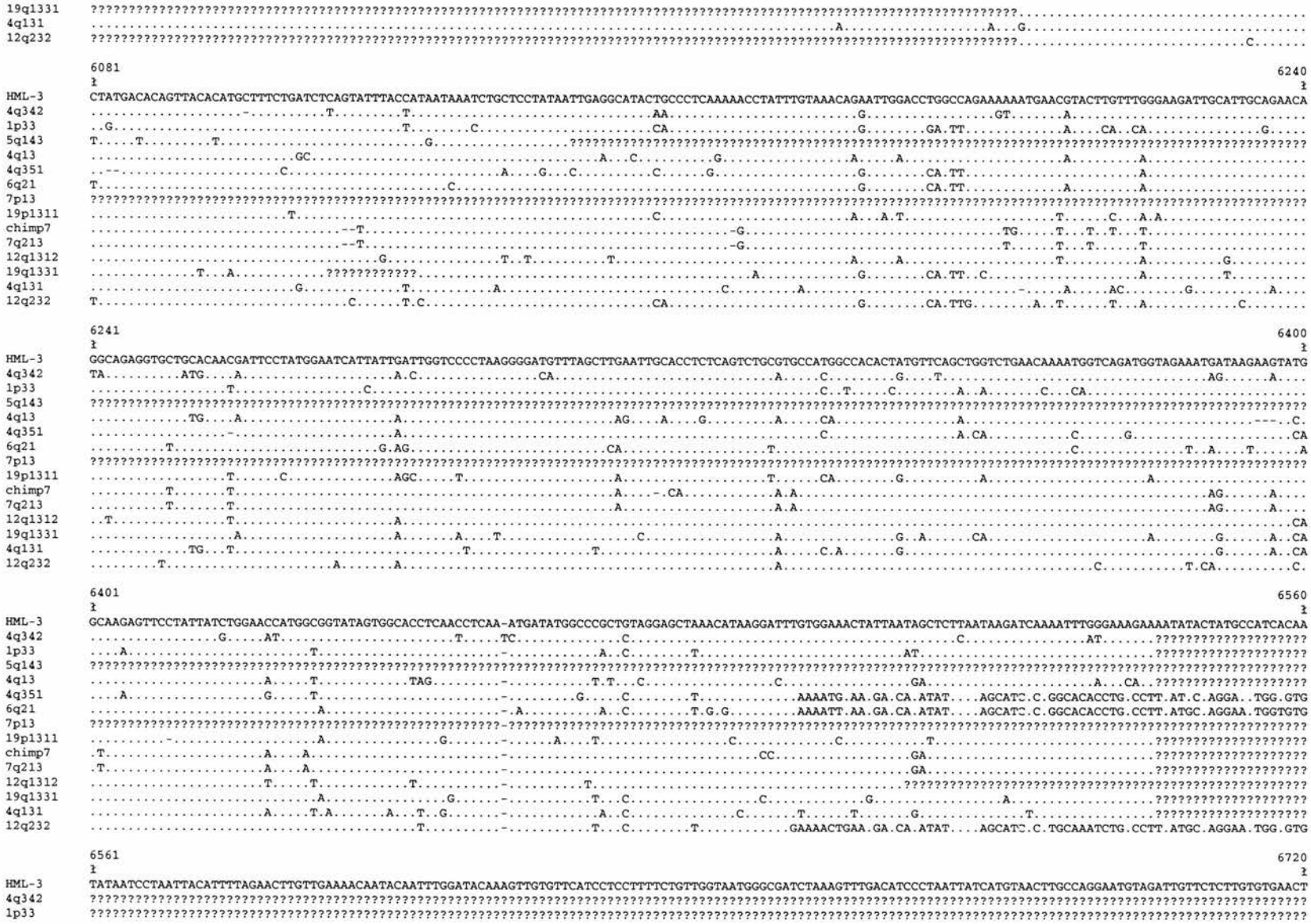



4q131
12q232
3841
4000
HML-3
4q342
1p33
5q143
4q13
4q351
6q21
7p13
19p1311
chimp7
7q213
12q1312
19q1331
4q131
12q232
4001
4160
HML-3
4q342
1p33
5q143
4q13
4q351
6q21
7p13
19p1311
chimp7
7q213
12q1312
19q1331
4q131
12q232
4161
4320
HML-3
4q342
1p33
5q143
4q13
4q351
6q21
7p13
19p1311
chimp7
7q213
12q1312
19q1331
4q131
12q232
4321
4480
HML-3
4q342
1p33
5q143

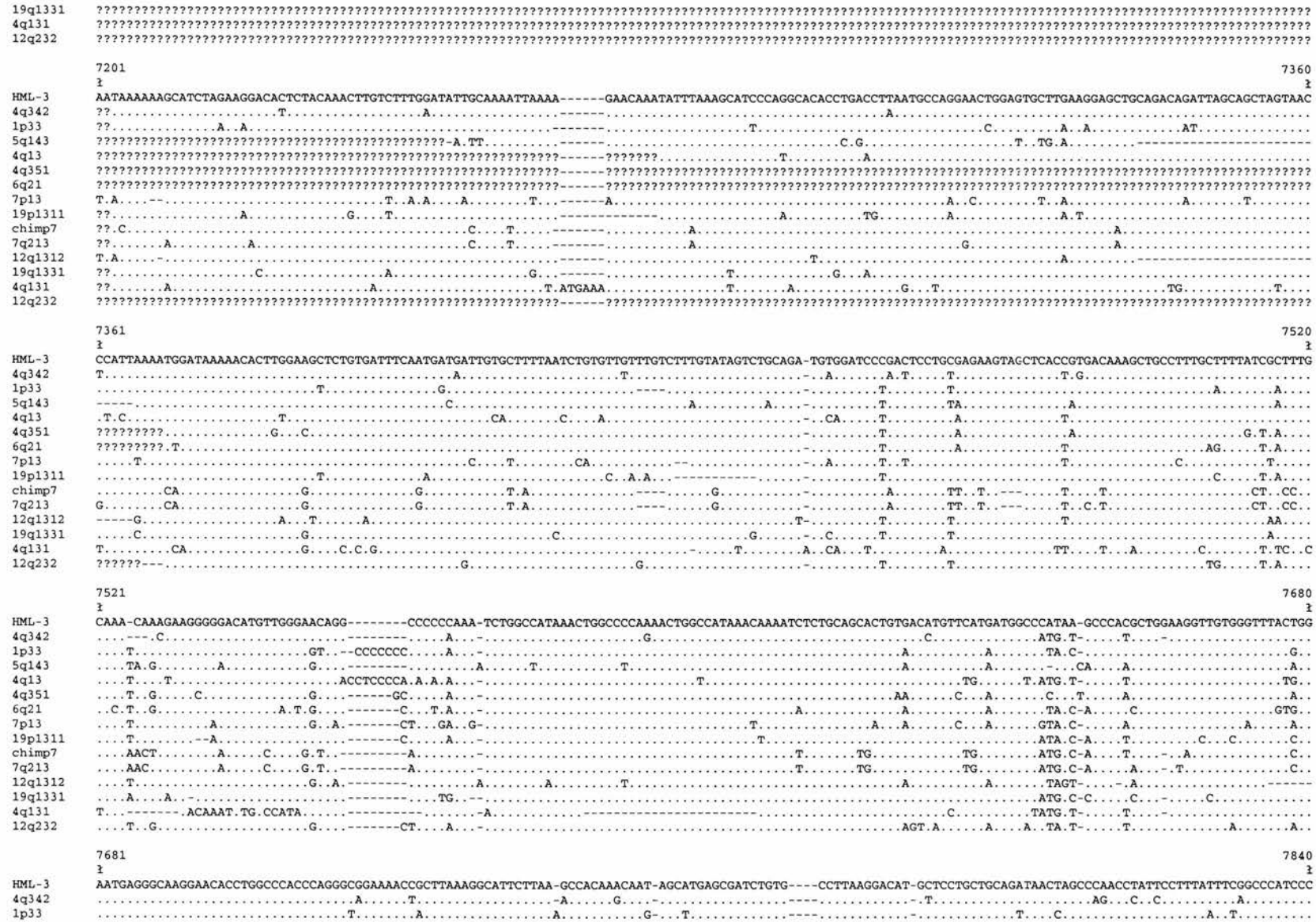

4q13
4q351
6q21
7p13
19p1311
chimp7
7q213
12q1312
19q1331
4q131
12q232
4481
4640
HML-3
4q342
1p33
5q143
4q13
4q351
6q21
7p13
19p1311
chimp7
7q213
12q1312
19q1331
4q131
12q232
4641
4800
HML-3
4q342
1p33
5q143
4q13
4q351
6q21
7p13
19p1311
chimp7
7q213
12q1312
19q1331
4q131
12q232
4801
4960
HML-3
4q342
1p33
5q143
4q13
4q351
6q21
7p13
19p1311
chimp7
7q213
12q1312
19q1331

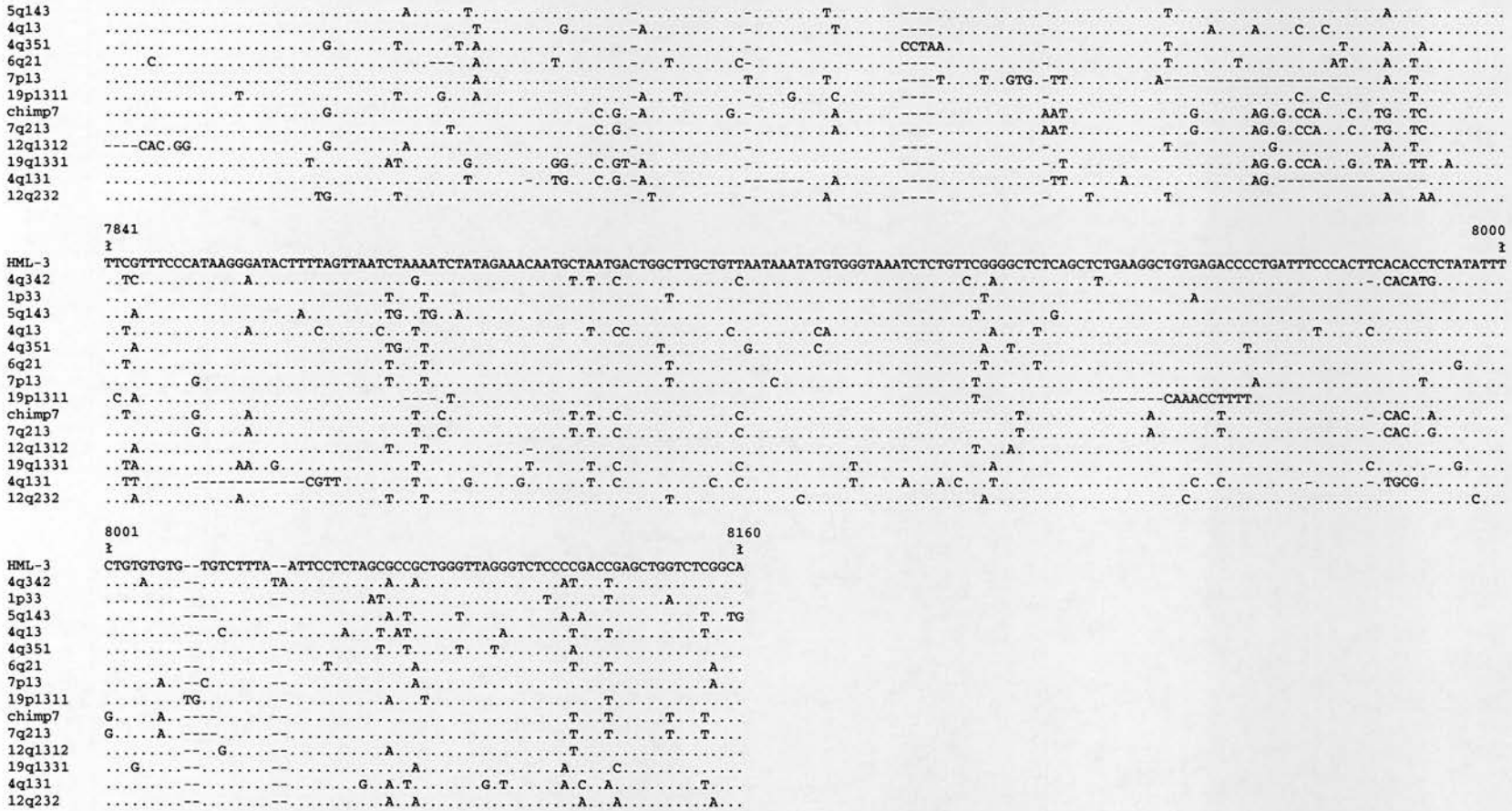

4q131
12q232
4961
5120
HML-3
4q342
1p33
5q143
4q13
4q351
6q21
7p13
19p1311
chimp7
7q213
12q1312
19q1331
4q131
12q232
5121
5280
HML-3
4q342
1p33
5q143
4q13
4q351
6q21
7p13
19p1311
chimp7
7q213
12q1312
19q1331
4q131
12q232
5281
5440
HML-3
4q342
1p33
5q143
4q13
4q351
6q21
7p13
19p1311
chimp7
7q213
12q1312
19q1331
4q131
12q232
5441
5600
HML-3
4q342
1p33
5q143


4q13
4q351
6q21
7p13
19p1311
chimp7
7q213
12q1312
19q1331
4q131
12q232
5601
5760
HML-3
4q342
1p33
5q143
4q13
4q351
6q21
7p13
19p1311
chimp7
7q213
12q1312
19q1331
4q131
12q232
5761
5920
HML-3
4q342
1p33
5q143
4q13
4q351
6q21
7p13
19p1311
chimp7
7q213
12q1312
19q1331
4q131
12q232
5921
6080
HML-3
4q342
1p33
5q143
4q13
4q351
6q21
7p13
19p1311
chimp7
7q213
12q1312

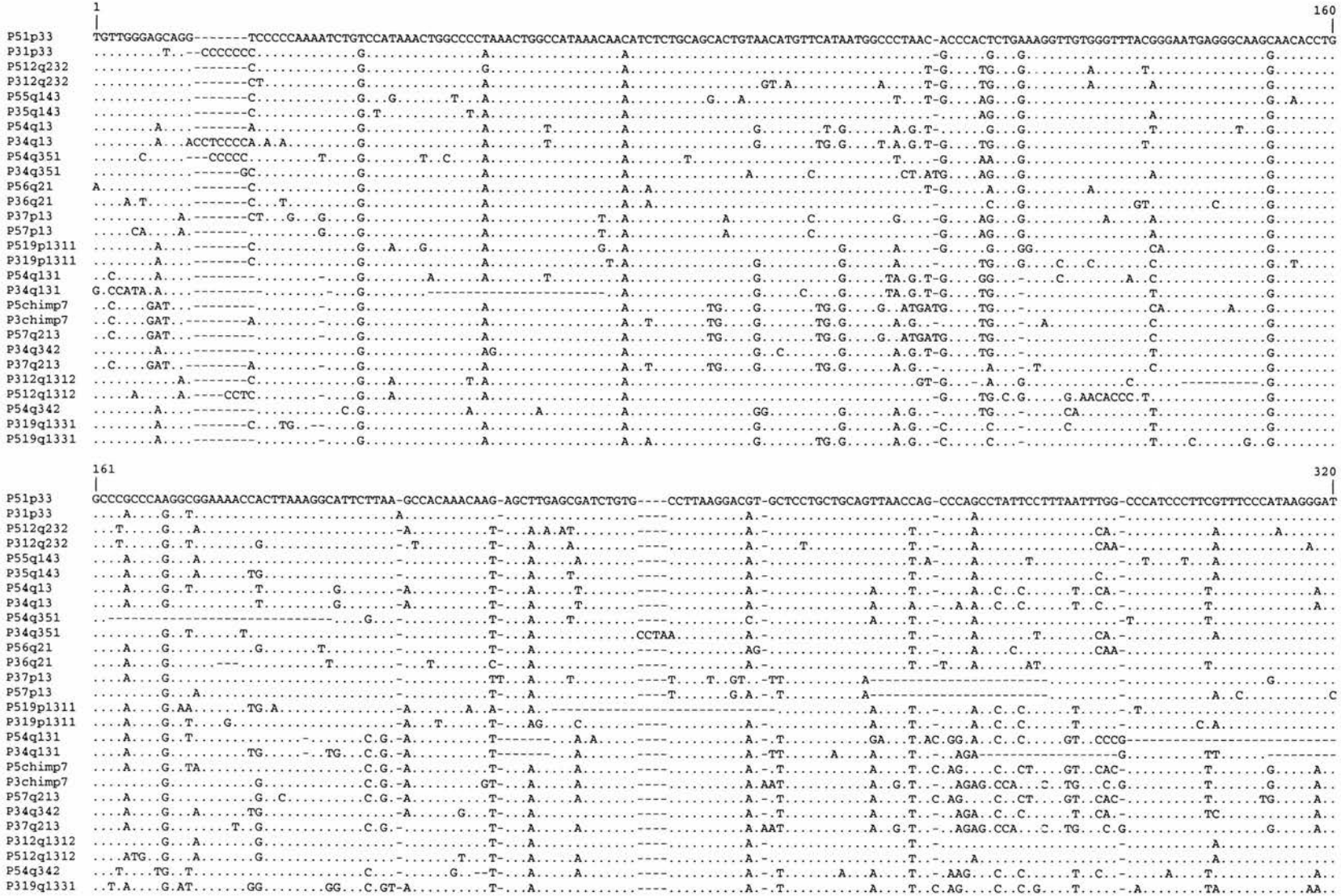


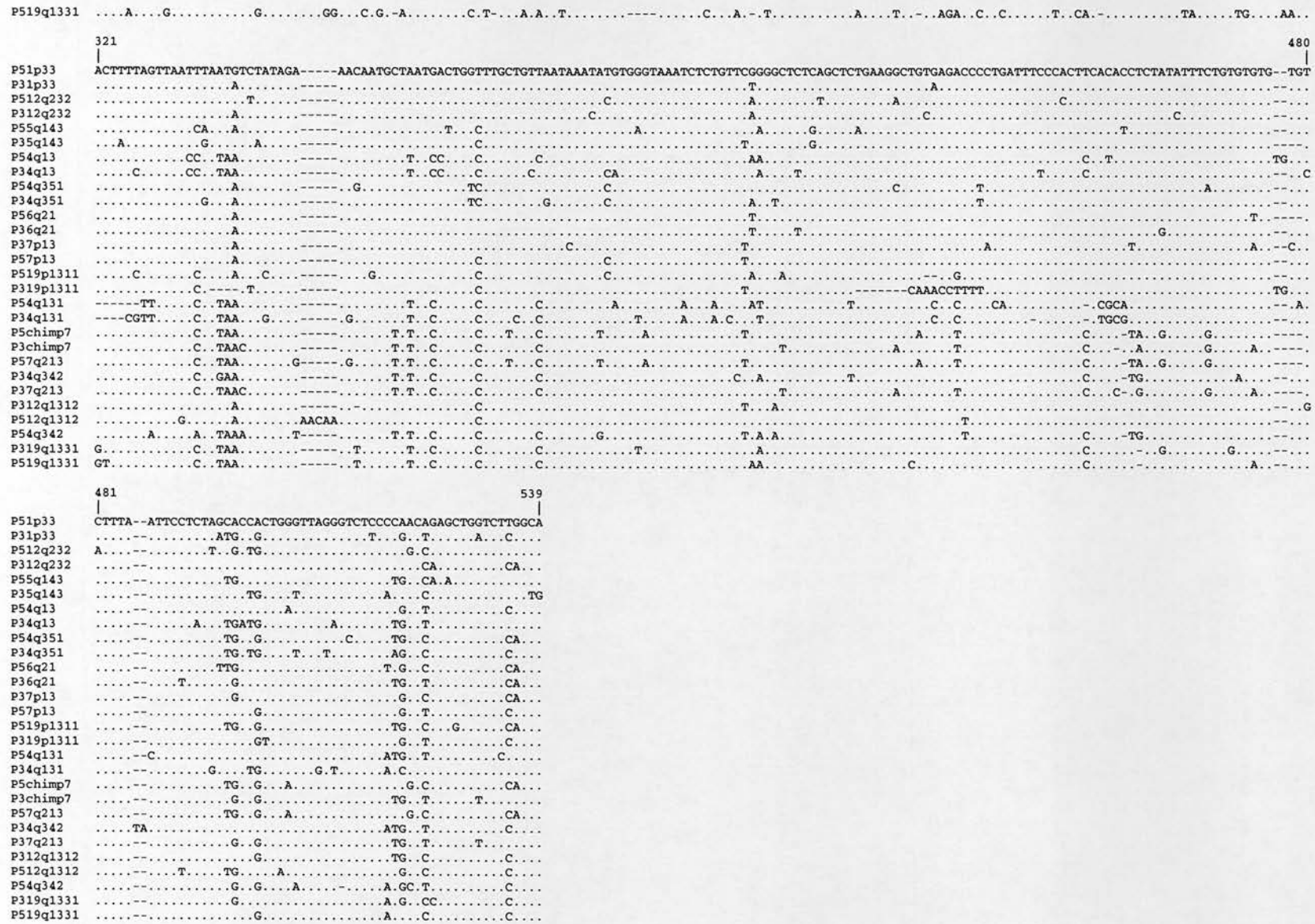

5q143
4q13
4q351
6q21
7p13
19p1311
chimp7
7q213
12q1312
19q1331
4q131
12q232
6721
6880
HML-3
4q342
1p33
5q143
4q13
4q351
6q21
7p13
19p1311
chimp7
7q213
12q1312
19q1331
4q131
12q232
6881
7040
HML-3
4q342
1p33
5q143
4q13
4q351
6q21
7p13
19p1311
chimp7
7q213
12q1312
19q1331
4q131
12q232
7041
7200
HML-3
4q342
1p33
5q143
4q13
4q351
6q21
7p13
19p1311
chimp7
7q213
12q1312


19q1331
4q131
12q232
7201
7360
HML-3
4q342
1p33
5q143
4q13
4q351
6q21
7p13
19p1311
chimp7
7q213
12q1312
19q1331
4q131
12q232
7361
7520
7521
7680
7681
7840

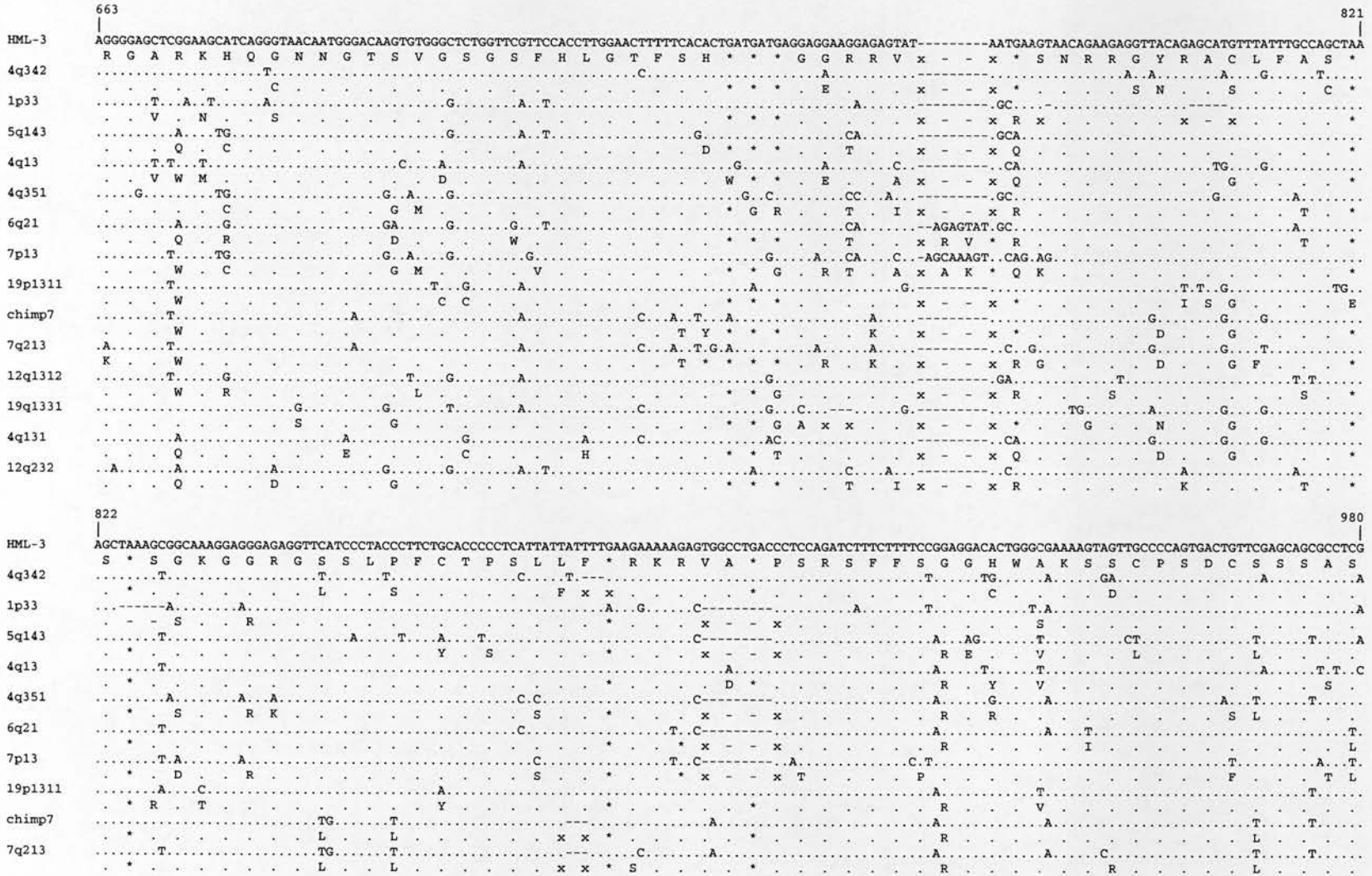

5q143
4q13
4q351
6q21
7p13
19p1311
chimp7
7q213
12q1312
19q1331
4q131
12q232
7841
8000
HML-3
4q342
1p33
5q143
4q13
4q351
6q21
7p13
19p1311
chimp7
7q213
12q1312
19q1331
4q131
12q232
8001
8160
HML-3
4q342
1p33
5q143
4q13
4q351
6q21
7p13
19p1311
chimp7
7q213
12q1312
19q1331
4q131
12q232

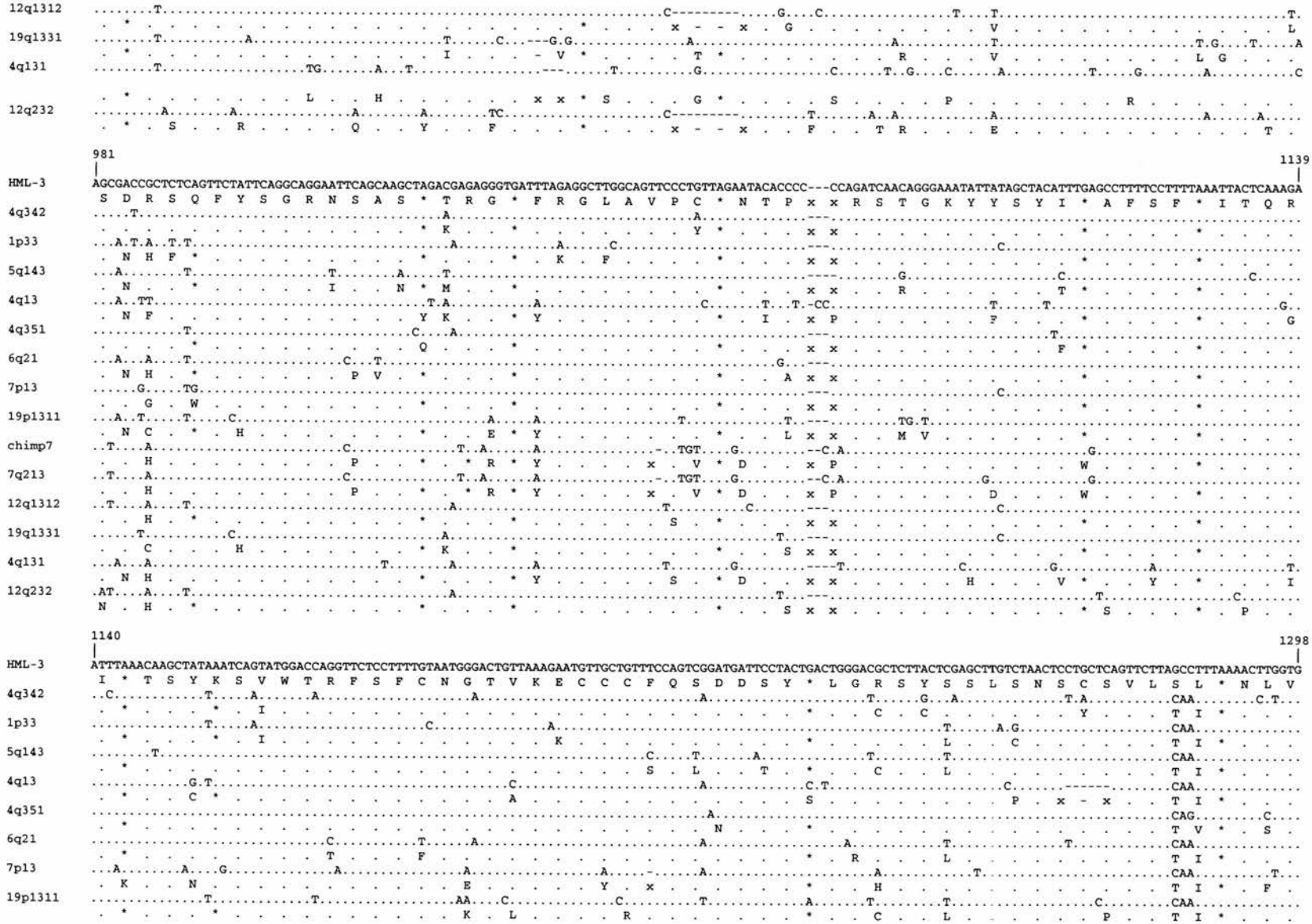
A.5 HERV-K(HML-3) Alignment of the respective LTRs of individual proviruses

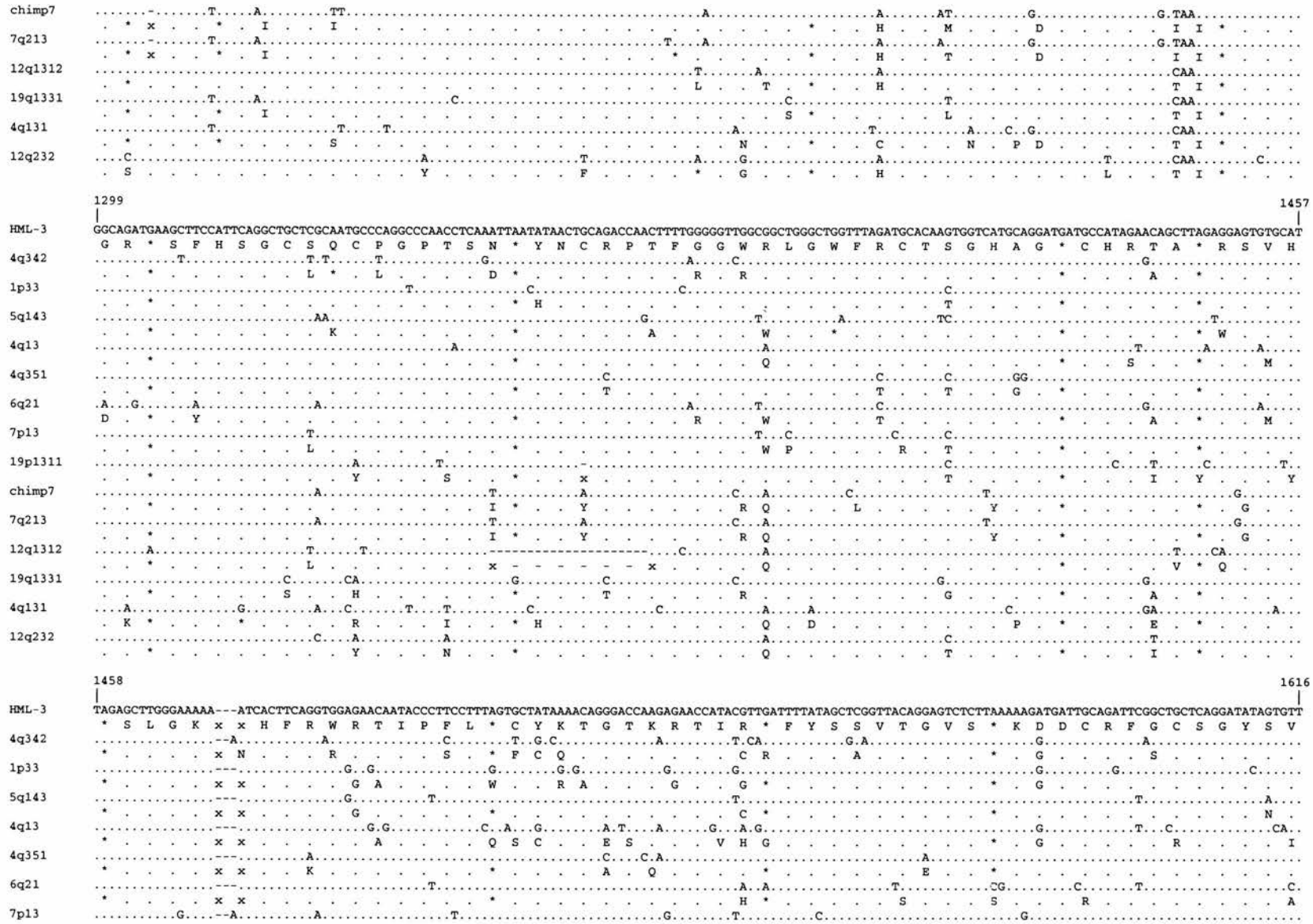


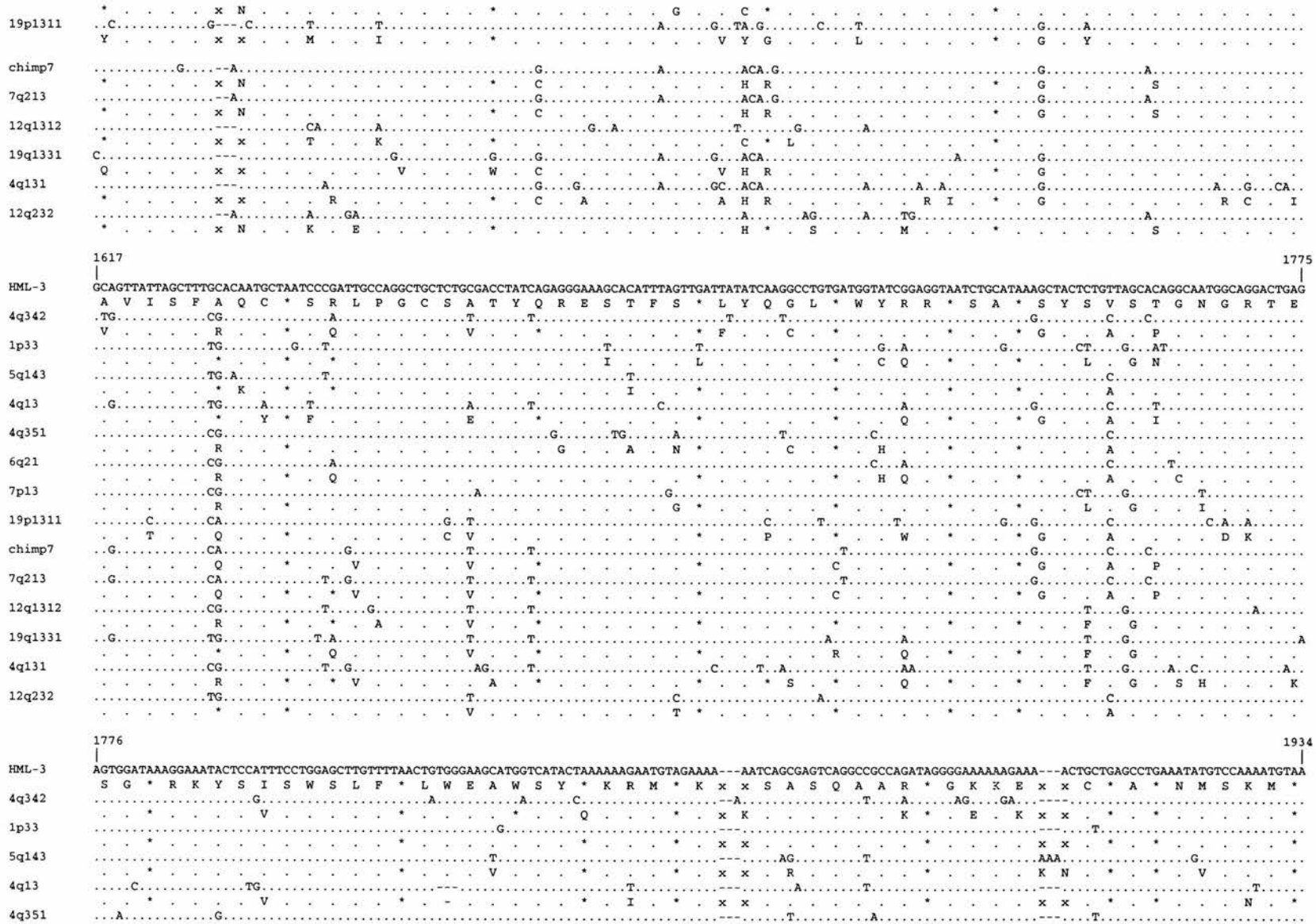


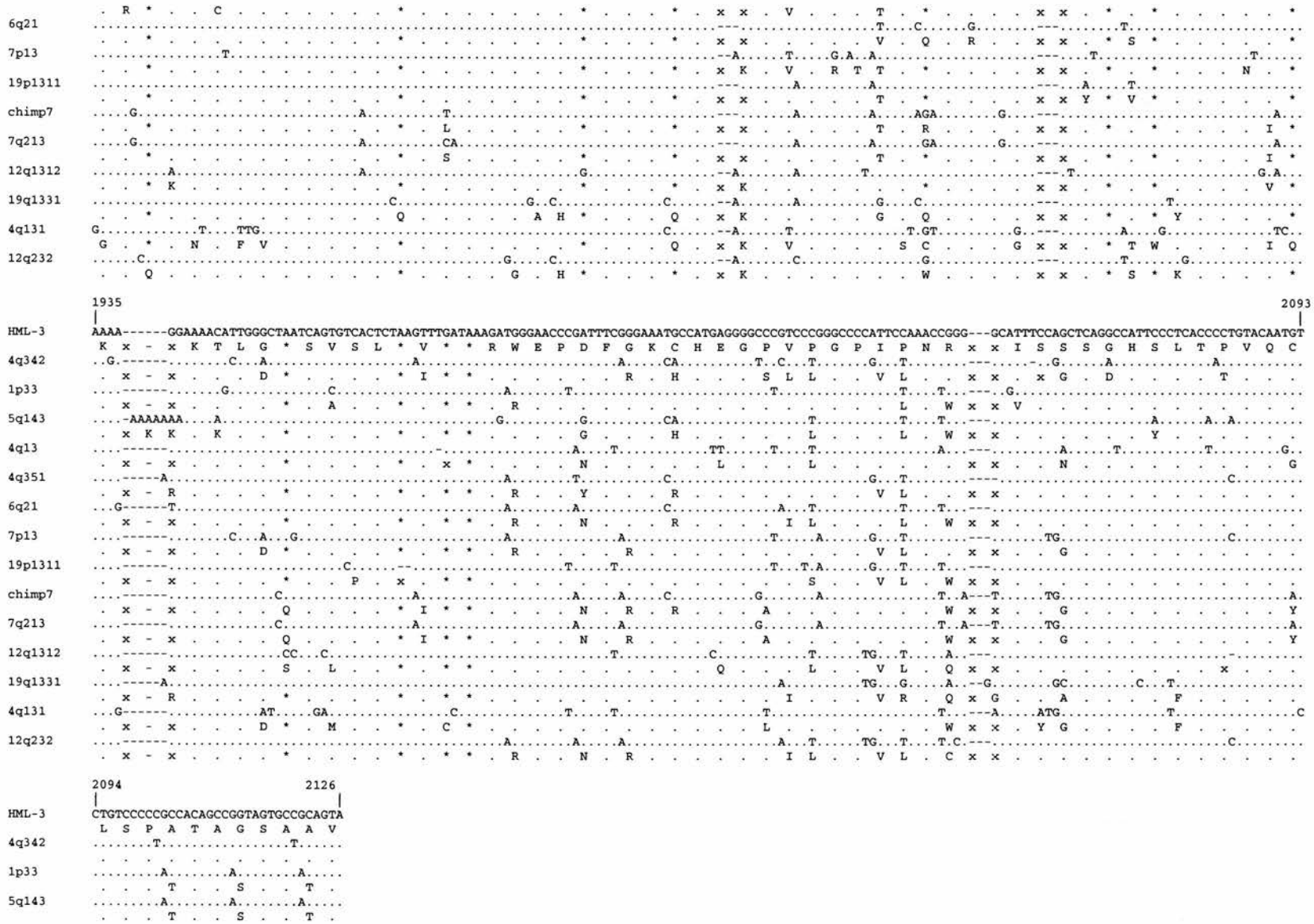
A.6 HERV-K(HML-3) Alignment of the Gag ORF

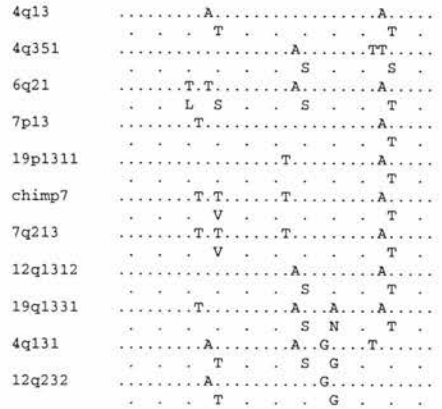



12q1312
19q1331
4q131
12q232
981
1139
HML-3
4q342
1p33
5q143
4q13
4q351
6q21
7p13
19p1311
chimp7
7q213
12q1312
19q1331
4q131
12q232
1140
1298
HML-3
4q342
1p33
5q143
4q13
4q351
6q21
7p13
19p1311

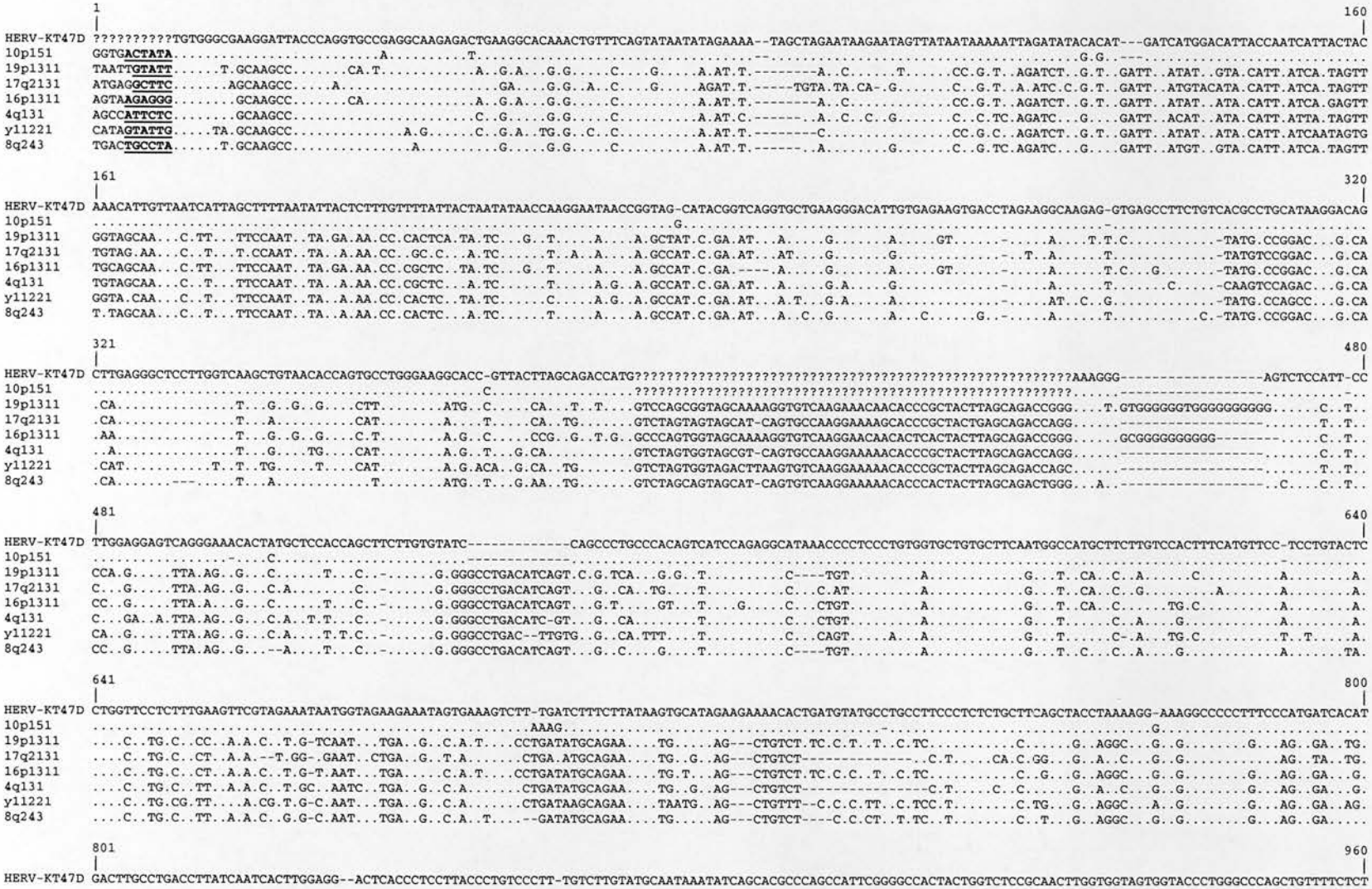

chimp7
7q213
12q1312
19q1331
4q131
12q232
1299
1457
HML-3
4q342
1p33
5q143
4q13
4q351
6q21
7p13
19p1311
chimp7
7q213
12q1312
19q1331
4q131
12q232
1458
1616
HML-3
4q342
1p33
5q143
4q13
4q351
6q21
7p13

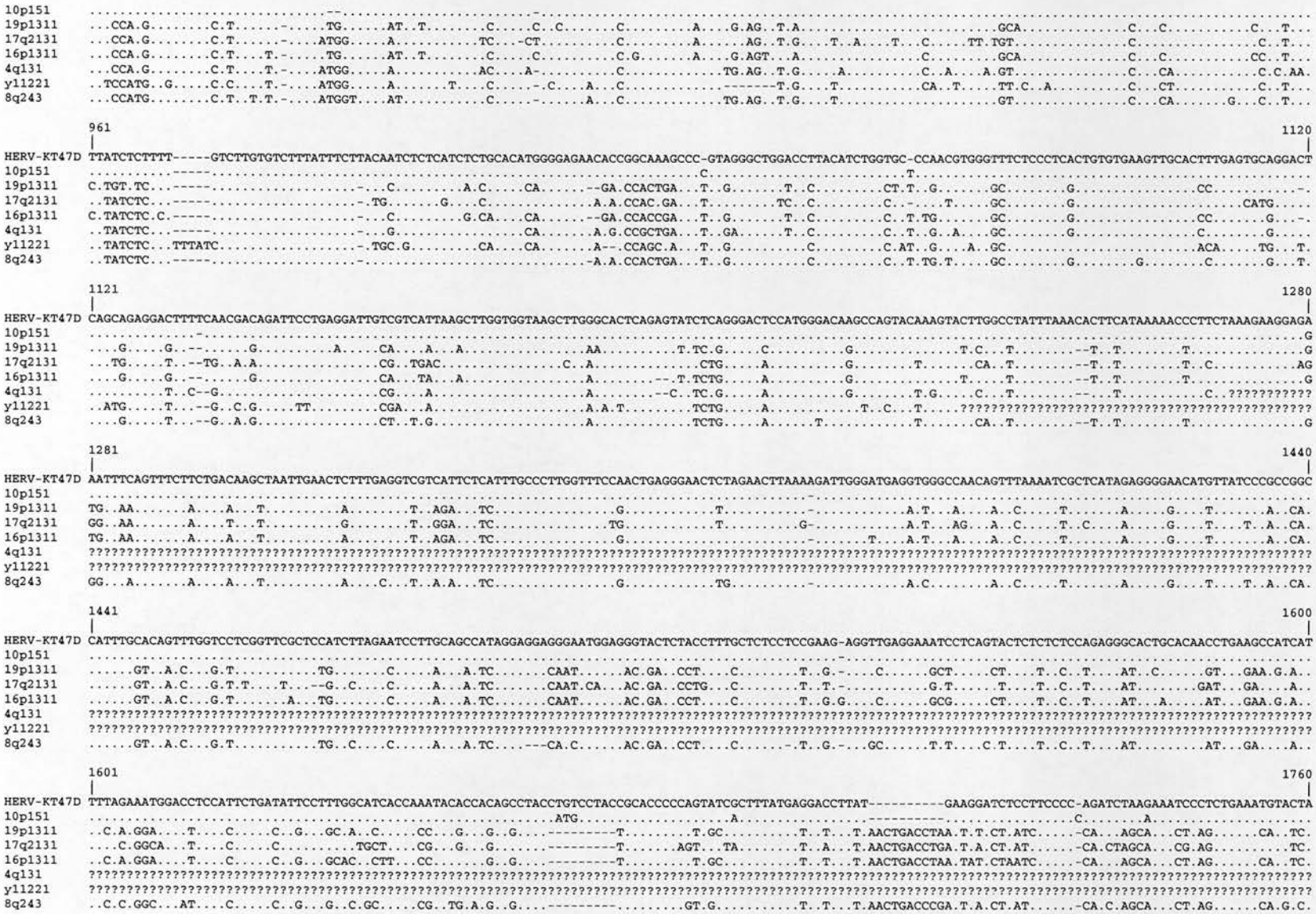

19p1311
chimp7
7q213
12q1312
19q1331
4q131
12q232
1617
1775
HML-3
GCAGTTATTAGCTTTGCACAATGCTAATCCCGATTGCCAGGCTGCTCTGCGACCTATCAGAGGGAAAGCACATTTAGTTGATTATATCAAGGCCTGTGATGGTATCGGAGGTAATCTGCATAAAGCTACTCTGTTAGCACAGGCAATGGCAGGACTGAG
4q342
1p33
5q143
4q13
4q351
6q21
7p13
1776
1934
AGTGGATAAAGGAAATACTCCATTTCCTGGAGCTTGTTTTAACTGTGGGAAGCATGGTCATACTAAAAAAGAATGTAGAAAA---AATCAGCGAGTCAGGCCGCCAGATAGGGGAAAAAAGAAA---ACTGCTGAGCCTGAAATATGTCCAAAATGTAA

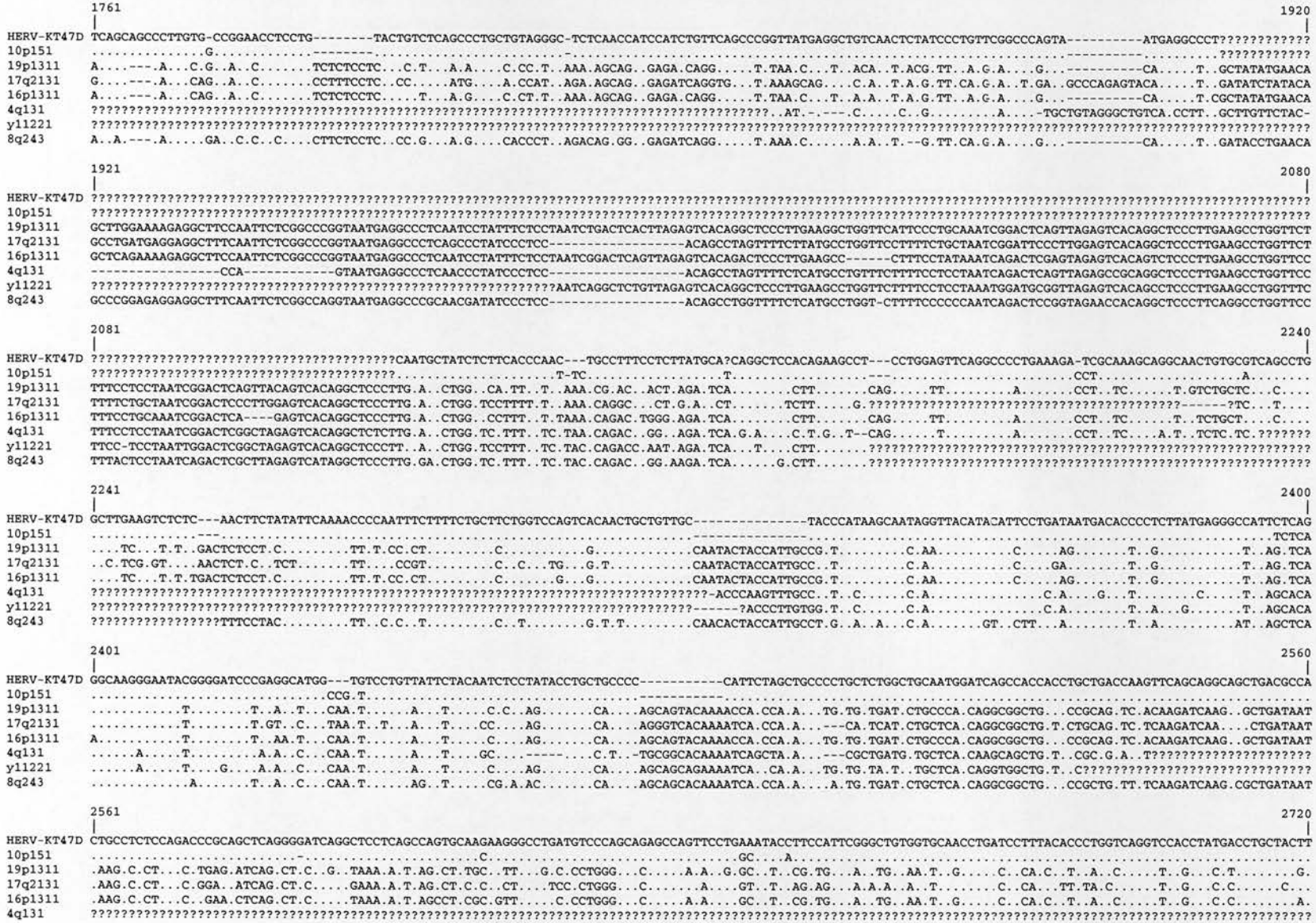

6q21
7p13
19p1311
chimp7
7q213
12q1312
19q1331
4q131
12q232
1935
2093
HML-3
AAAA------GGAAAACATTGGGCTAATCAGTGTCACTCTAAGTTTGATAAAGATGGGAACCCGATTTCGGGAAATGCCATGAGGGGCCCGTCCCGGGCCCCATTCCAAACCGGG---GCATTTCCAGCTCAGGCCATTCCCTCACCCCTGTACAATGT
4q342
1p33
5q143
4q13
4q351
6q21
7p13
19p1311
chimp7
7q213
12q1312
19q1331
4q131
12q232
2094
2126
CTGTCCCCCGCCACAGCCGGTAGTGCCGCAGTA
4q342
1p33
5q143

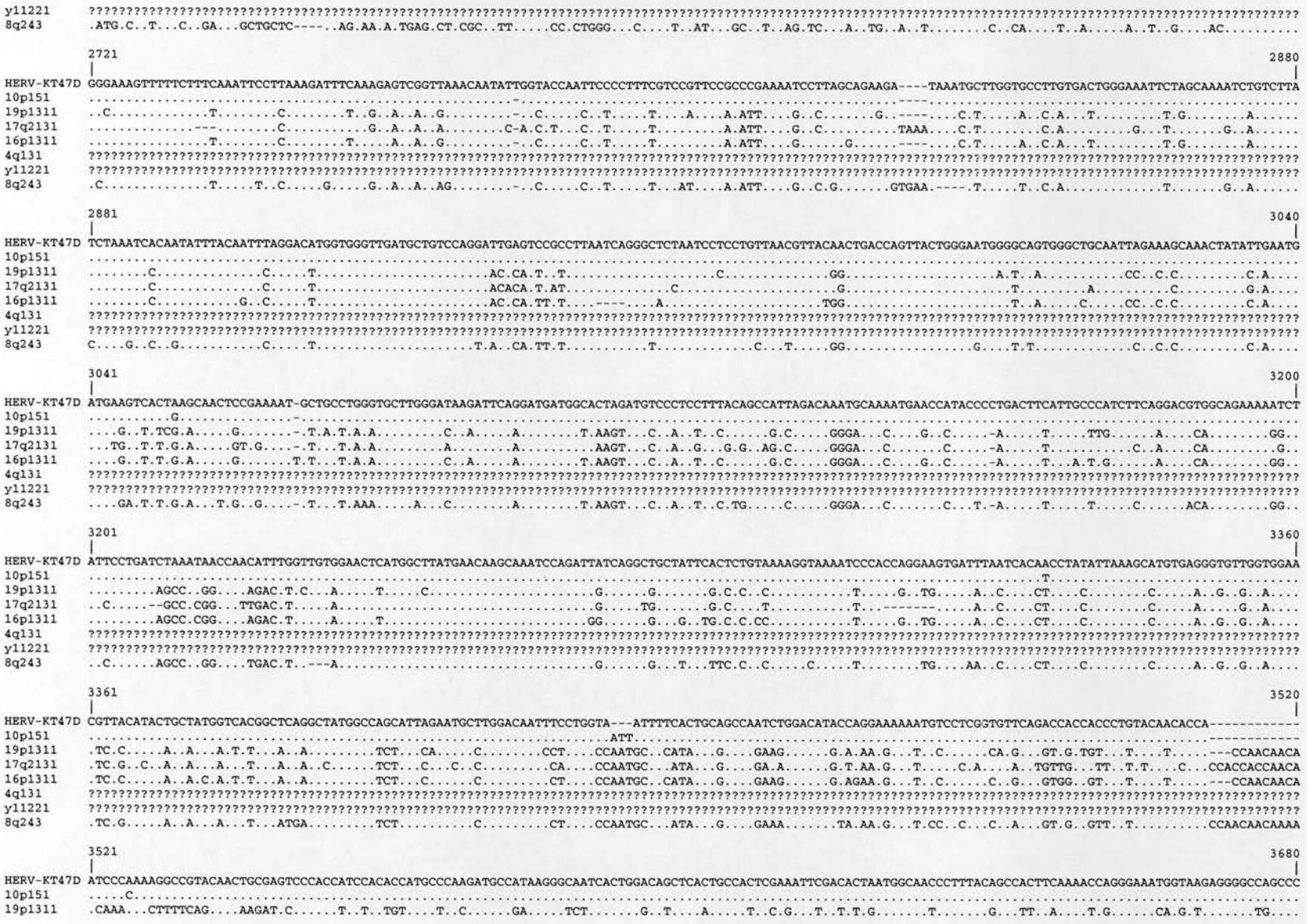

4q13
4q351
6q21
7p13
19p1311
chimp7
7q213
12q1312
19q1331
4q131
12q232

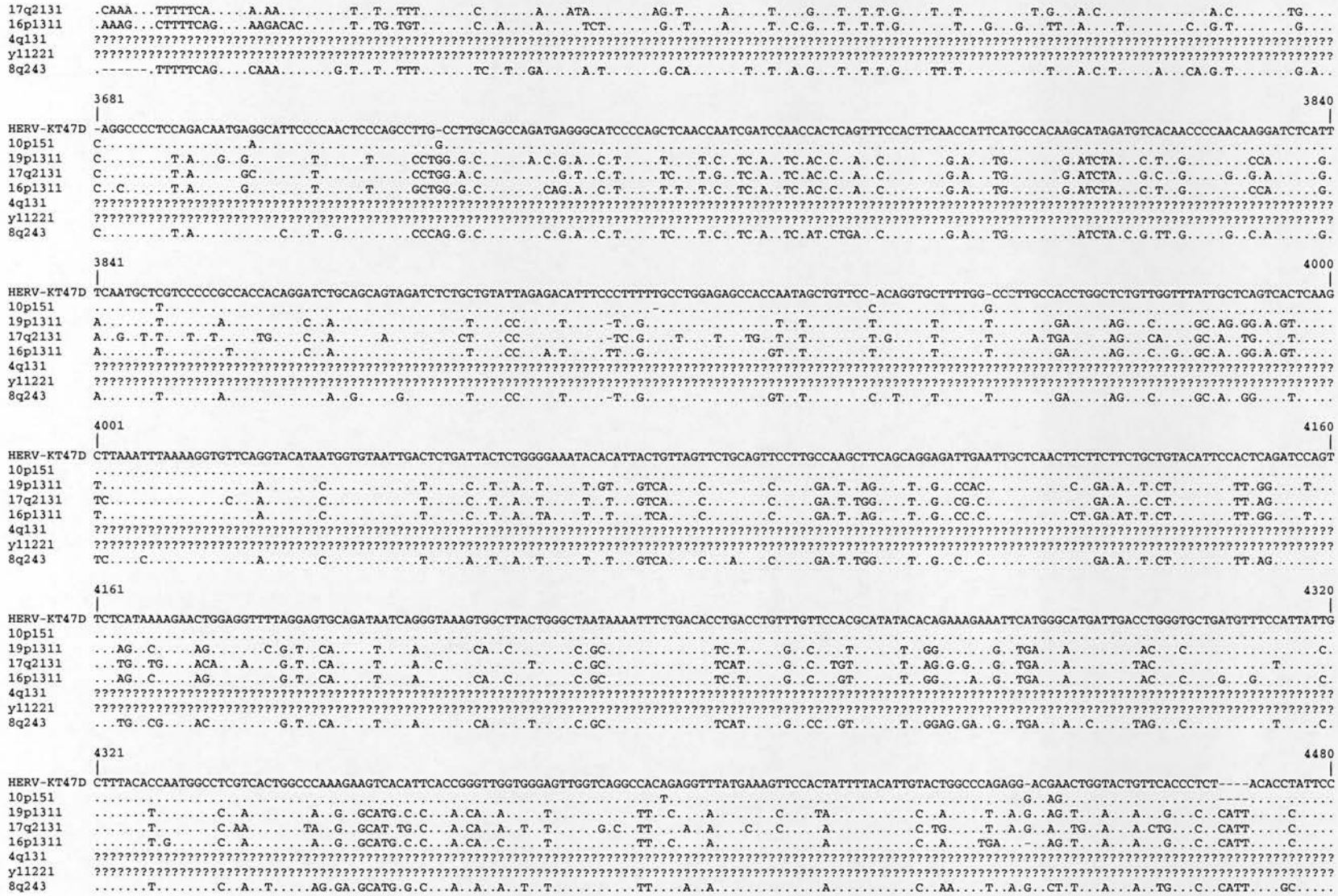
A.7 HERV-K(HML-4) Alignment of complete proviruses used in study

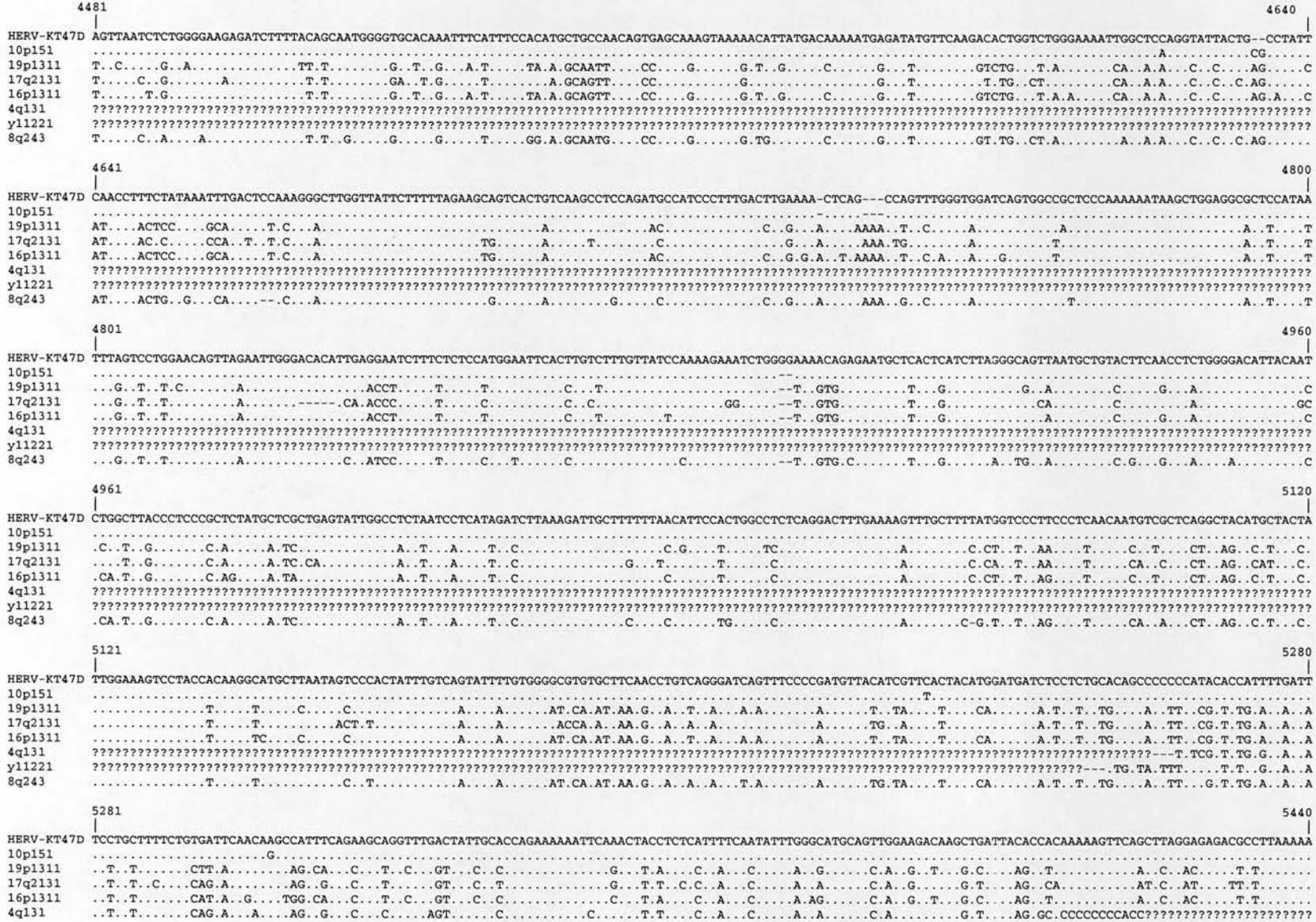


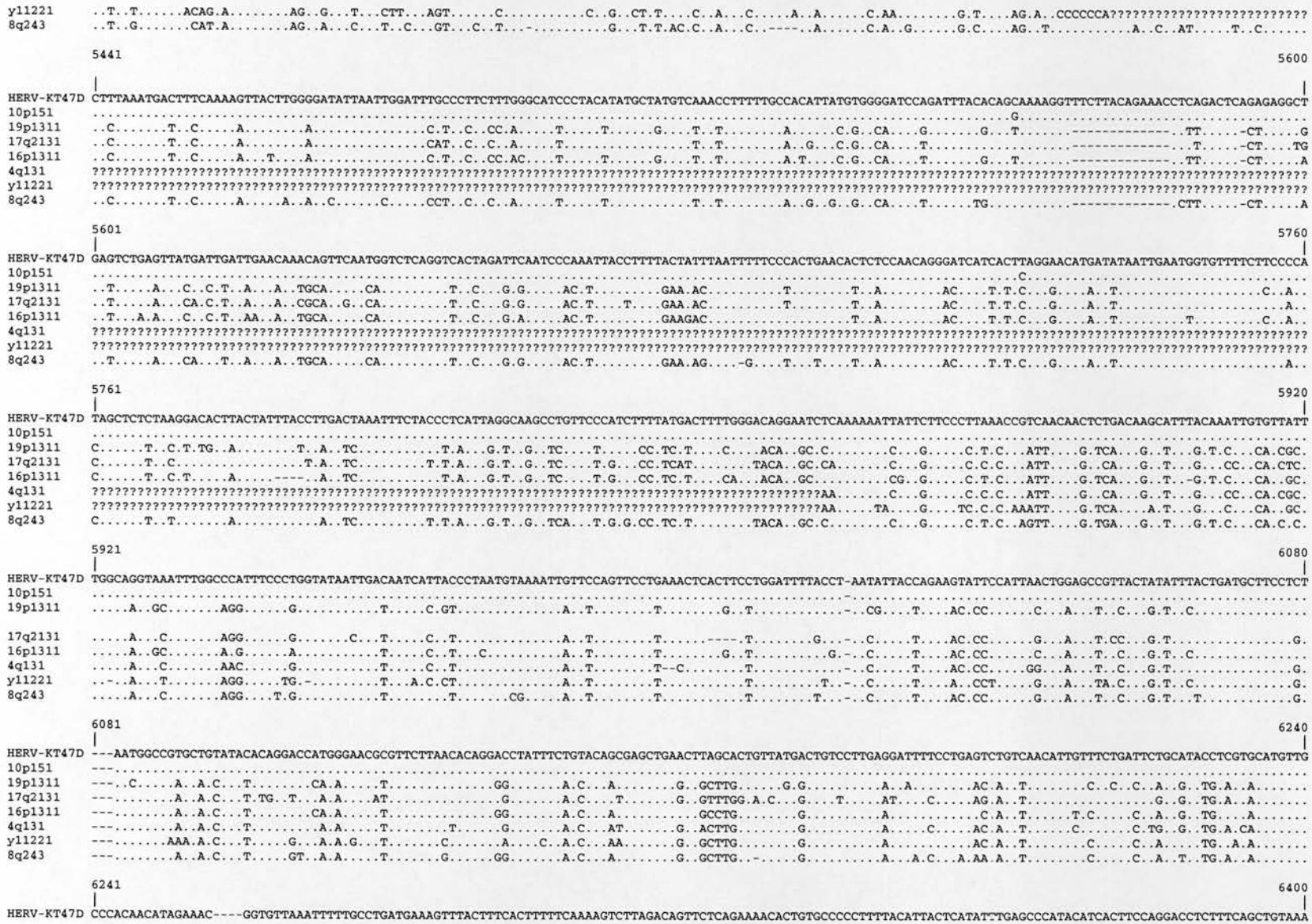


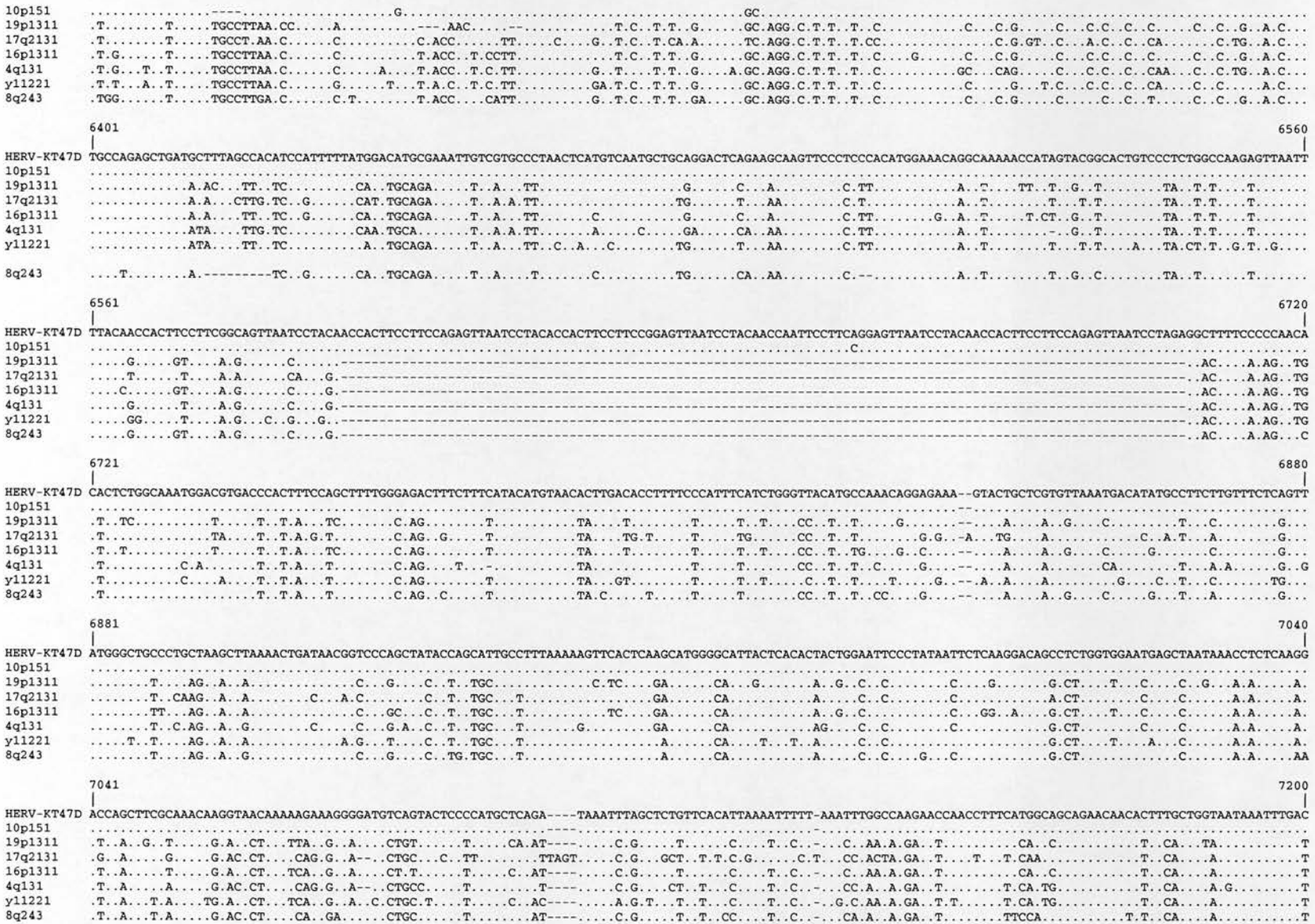


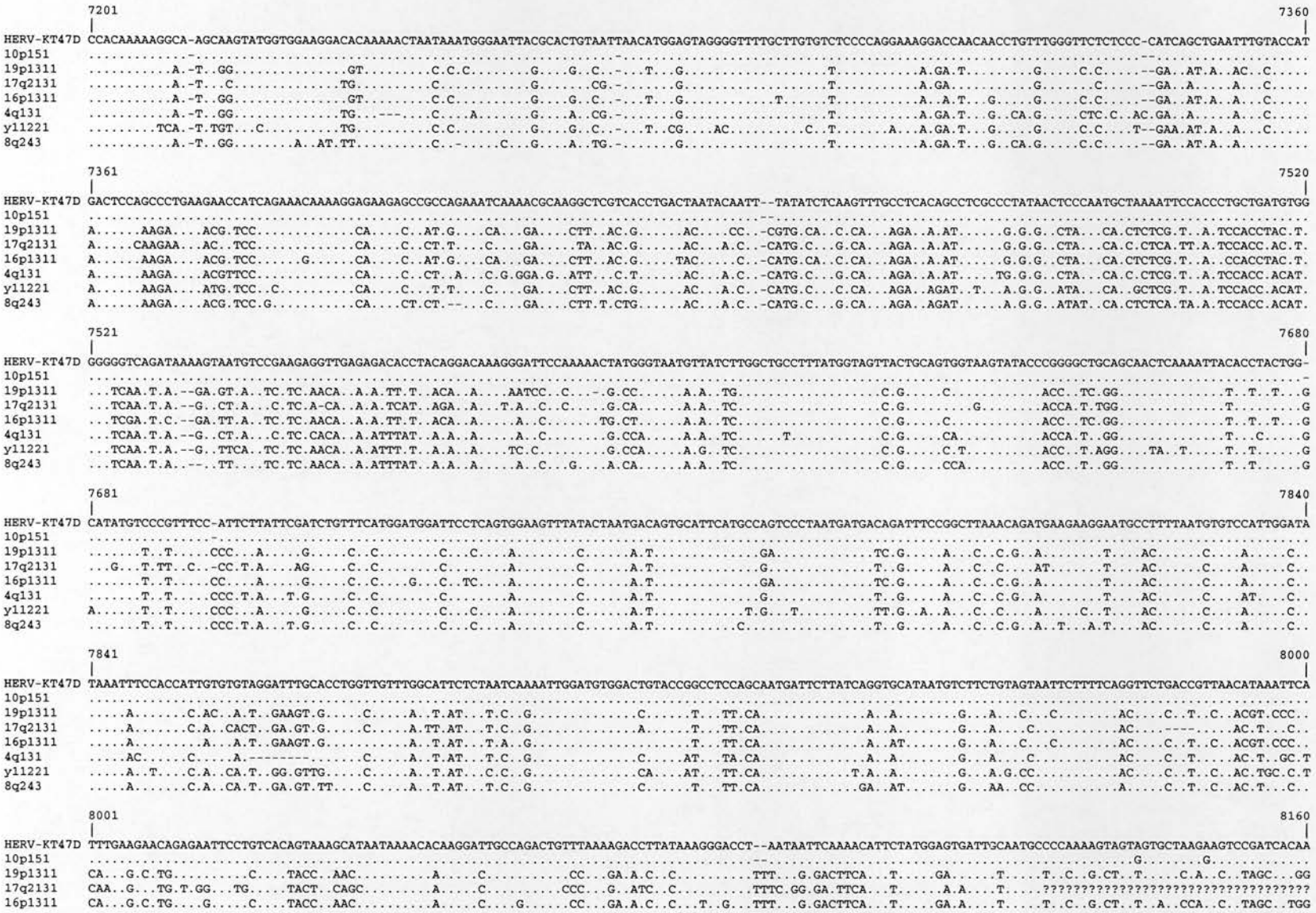





4481
4640
HERV-KT47D AGTTAATCTCTGGGGAAGAGATCTTTTACAGCAATGGGGTGCACAAATTTCATTTCCACATGCTGCCAACAGTGAGCAAAGTAAAAACATTATGACAAAAATGAGATATGTTCAAGACACTGGTCTGGGAAAATTGGCTCCAGGTATTACTG--CCTATT
10p151
19p1311
17q2131
16p1311
4q131
y11221
8q243
4641
4800
HERV-KT47D CAACCTTTCTATAAATTTGACTCCAAAGGGCTTGGTTATTCTTTTTAGAAGCAGTCACTGTCAAGCCTCCAGATGCCATCCCTTTGACTTGAAAA-CTCAG---CCAGTTTGGGTGGATCAGTGGCCGCTCCCAAAAAATAAGCTGGAGGCGCTCCATAA
4801
4960
HERV-KT47D TTTAGTCCTGGAACAGTTAGAATTGGGACACATTGAGGAATCTTTCTCTCCATGGAATTCACTTGTCTTTGTTATCCAAAAGAAATCTGGGGAAAACAGAGAATGCTCACTCATCTTAGGGCAGTTAATGCTGTACTTCAACCTCTGGGGACATTACAAT
4961
5120
HERV-KT47D CTGGCTTACCCTCCCGCTCTATGCTCGCTGAGTATTGGCCTCTAATCCTCATAGATCTTAAAGATTGCTTTTTTAACATTCCACTGGCCTCTCAGGACTTTGAAAAGTTTGCTTTTATGGTCCCTTCCCTCAACAATGTCGCTCAGGCTACATGCTACTA
5121
5280
HERV-KT47D TTGGAAAGTCCTACCACAAGGCATGCTTAATAGTCCCACTATTTGTCAGTATTTTGTGGGGCGTGTGCTTCAACCTGTCAGGGATCAGTTTCCCCGATGTTACATCGTTCACTACATGGATGATCTCCTCTGCACAGCCCCCCCATACACCATTTTGATT
5281
5440
HERV-KT47D TCCTGCTTTTCTGTGATTCAACAAGCCATTTCAGAAGCAGGTTTGACTATTGCACCAGAAAAAATTCAAACTACCTCTCATTTTCAATATTTGGGCATGCAGTTGGAAGACAAGCTGATTACACCACAAAAAGTTCAGCTTAGGAGAGACGCCTTAAAAA







7201 7360
HERV-KT47D CCACAAAAAGGCA-AGCAAGTATGGTGGAAGGACACAAAAACTAATAAATGGGAATTACGCACTGTAATTAACATGGAGTAGGGGTTTTGCTTGTGTCTCCCCAGGAAAGGACCAACAACCTGTTTGGGTTCTCTCCC-CATCAGCTGAATTTGTACCAT
10p151 .....-.....
19p1311 .....A.-T..GG.....GT.....C.C.C.....G....G..C..-...T..G.....T.....A.GA.T.....G.....C.C.....--GA..AT.A..AC..C.....
17q2131 .....A.-T...C.....TG.....C.....G.....CG.-.....G.....T.....A.GA.....G.....C.....--GA..A.....A...C.....
16p1311 .....A.-T..GG.....GT.....C.C.....G....G..C..-...T...G.....T.....T.....A..A.T...G.....G.....C.C.....--GA..AT.A..A...C.....
4q131 .....A.-T..GG.....TG...---....C....A.....G....A..CG.-.....G.....T.....A.GA.T...G..CA.G.....CTC.C..AC.GA..A.....A...C.....
y11221 .....TCA.-T.TGT...C.....TG.....C.C.....G....G..C..-...T..CG....AC.....C..T.....A...A.GA.T...G.....G.....C.C....T--GAA.AT.A..A...C.....
8q243 .....A.-T..GG.....A..AT.TT.....C..-.....C...G....A..TG.-.....G.....T.....A.GA.T...G..CA.G.....C.C.....--GA..AT.A..A.....

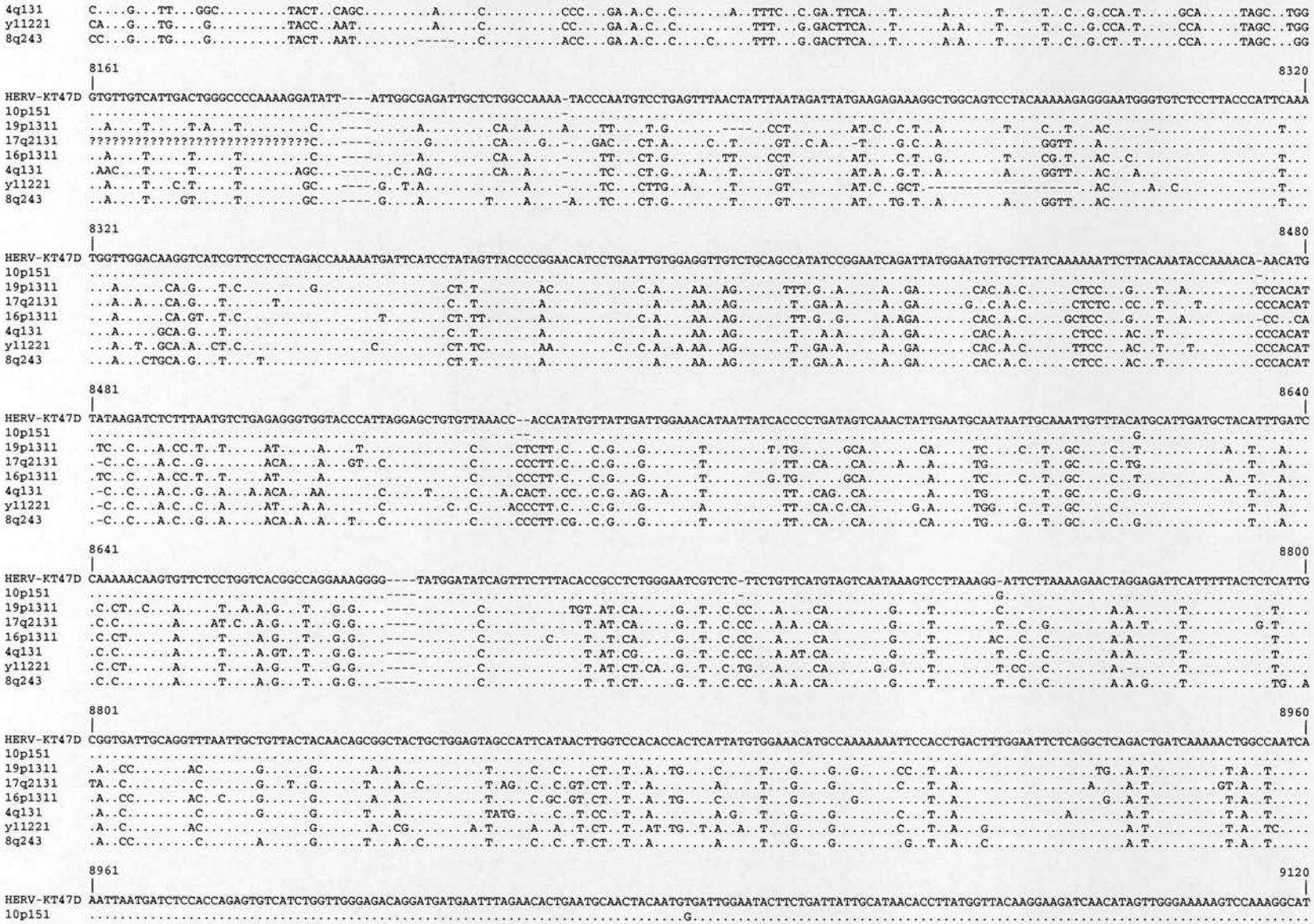
7361 7520
HERV-KT47D GACTCCAGCCCTGAAGAACCATCAGAAACAAAAGGAGAAGAGCCGCCAGAAATCAAAACGCAAGGCTCGTCACCTGACTAATACAATT--TATATCTCAAGTTTGCCTCACAGCCTCGCCCTATAACTCCCAATGCTAAAATTCCACCCTGCTGATGTGG
10p151 .....--.....
19p1311 A.....AAGA....ACG.TCC.....CA....C..AT.G....CA...GA....CTT..AC.G.....AC....CC..-CGTG.CA..C.CA...AGA..A.AT.....G.G.G..CTA...CA.CTCTCG.T..A.TCCACCTAC.T.
17q2131 A.....CAAGAA...AC..TCC.....CA....C..CT.T....C....GA.....TA..AC.G.....AC...A.C..-CATG.C...G.CA...AGA..A.AT.....G.G.G..CTA...CA.C.CTCA.TT.A.TCCACC.AC.T.
16p1311 A.....AAGA....ACG.TCC.....G.....CA....C..AT.G....CA...GA....CTT..AC.G.....TAC.....C..-CATG.CA..C.CA...AGA..A.AT.....G.G.G..CTA...CA.CTCTCG.T..A..CCACCTAC.T.
4q131 A.....AAGA....ACGTTCC.....CA....C..CT..A...C.G.GGA.G..ATT...C.T.....AC...A.C..-CATG.C...G.CA...AGA..A.AT.....TG.G.G..CTA...CA.C.CTCG.T..A.TCCACC.ACAT.
y11221 A.....AAGA....ATG.TCC..C.....CA....C...T.T....C....GA....CTT..AC.G.....AC...A.C..-CATG.C...C.CA...AGA..AGAT..T...A.G.G..ATA...CA..GCTCG.T..A.TCCACC.ACAT.
8q243 A.....AAGA....ACG.TCC.G.....CA....CT.CT.--..C....GA....CTT.T.CTG.....AC...A.C..-CATG.C...G.CA...AGA..AGAT.....A.G.G..ATAT..CA.CTCTCA.TA.A.TCCACC.ACAT.

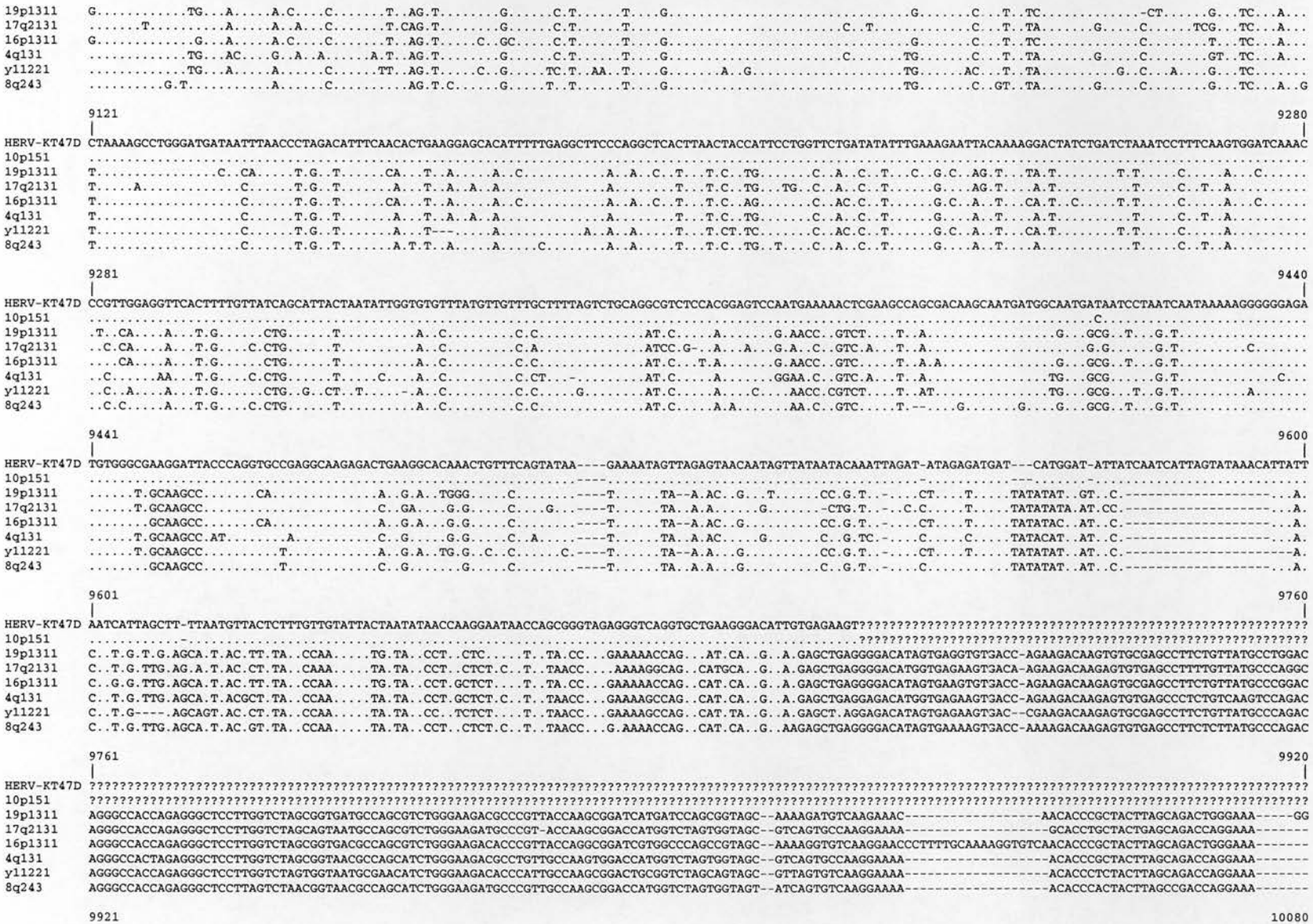
7521 7680
HERV-KT47D GGGGGTCAGATAAAAGTAATGTCCGAAGAGGTTGAGAGACACCTACAGGACAAAGGGATTCCAAAAACTATGGGTAATGTTATCTTGGCTGCCTTTATGGTAGTTACTGCAGTGGTAAGTATACCCGGGGCTGCAGCAACTCAAAATTACACCTACTGG-
10p151 .....-
19p1311 ...TCAA.T.A.--GA.GT.A..TC.TC.AACA..A.A.TT.T..ACA..A....AATCC..C...-.G.CC.....A.A..TG.....C.G.....C.....ACC..TC.GG.....T..T..T...G
17q2131 ...TCAA.T.A.--G..CT.A...C.TC.A-CA..A.A.TCAT..AGA..A...T.A..C..C.....G.CA.....A.A..TC.....C.G.....G.....ACCA.T.TGG.....T.....G
16p1311 ...TCGA.T.C.--GA.TT.A..TC.TC.AACA..A.A.TT.T..ACA..A.....A..C.....TG.CT.....A.A..TC.....C.G.....C.....ACC..TC.GG.....T..T..T...G
4q131 ...TCAA.T.A.--G..CT.A...C.TC.CACA..A.ATTTAT..A.A..A.....A..C.....G.CCA.....A.A..TC.....T.....C.G.....CA.....ACCA.T..GG.....T...C.....G
y11221 ...TCAA.T.A.--G..TTCA..TC.TC.AACA..A.ATTT.T..A.A..A....TC.C.....G.CCA.....A.G..TC.....C.G.....C.T.....ACC..T.AGG....TA..T.....T..T.....G
8q243 ...TCAA.T.A..--..TT....TC.TC.AACA..A.ATTTAT..A.A..A.....A..C...G....A.CA.....A.A..TC.....C.G.....CCA.....ACC..T..GG.....T..T.....G

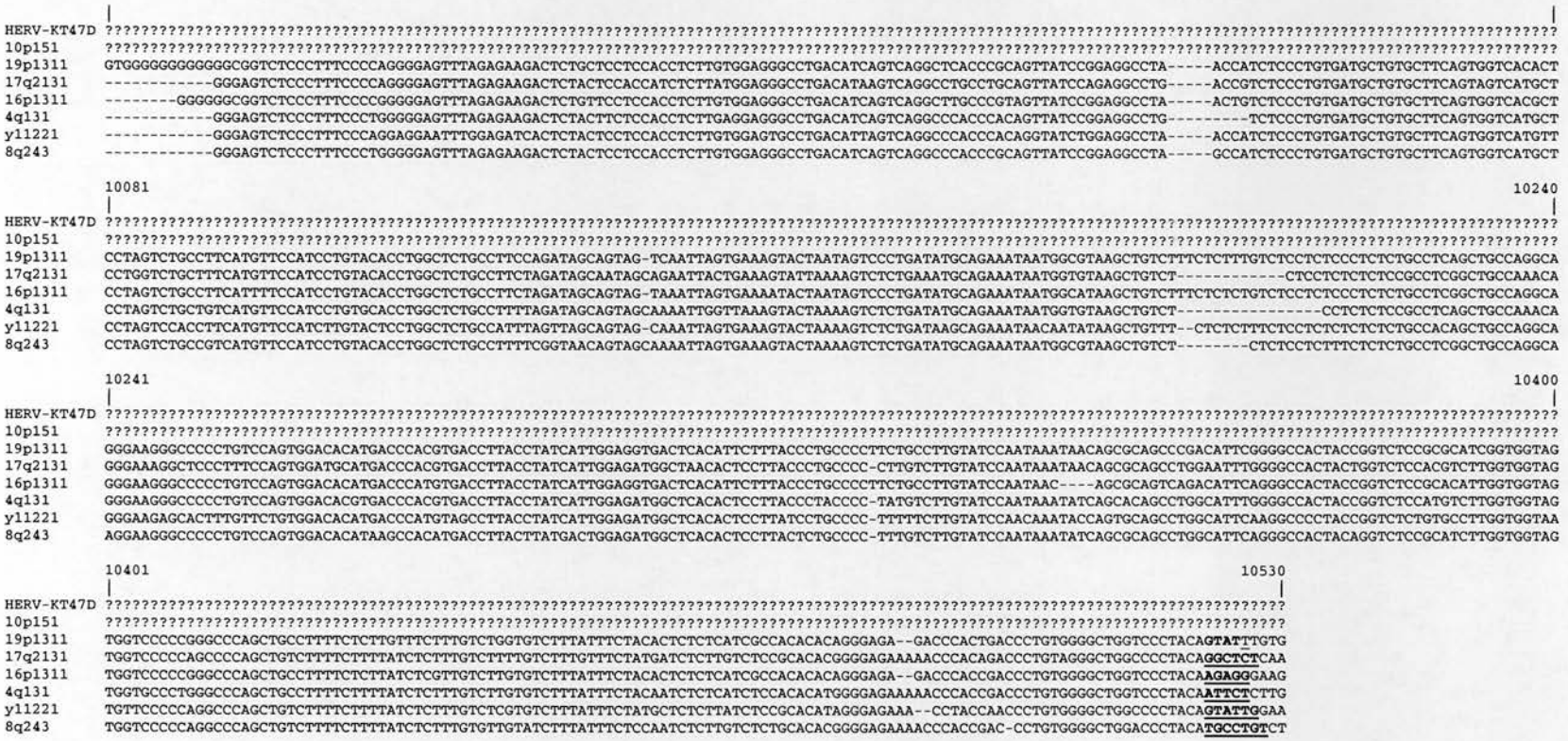
7681 7840
HERV-KT47D CATATGTCCCGTTTCC-ATTCTTATTCGATCTGTTTCATGGATGGATTCCTCAGTGGAAGTTTATACTAATGACAGTGCATTCATGCCAGTCCCTAATGATGACAGATTTCCGGCTTAAACAGATGAAGAAGGAATGCCTTTTAATGTGTCCATTGGATA
10p151 .....-
19p1311 .....T..T.....CCC...A.....G.....C..C.....C...C....A.....C.....A.T.....GA.....TC.G.....A...C..C.G..A.....T....AC.....C....A.....C..
17q2131 ...G...T.TT..C..-CC.T.A....AG.....C..C.....C.....A.....C.....A.T.....G.....T..G.....A...C..C....AT.....T....AC.....C....A.....C..
16p1311 .....T..T.....CC....A.....G.....C..C....G...C..TC....A.....C.....A.T.....GA.....TC.G.....A...C..C.G..A.....T....AC.....C....A.....C..
4q131 .....T..T.....CCC.T.A...T.G.....C..C.....C.....A.....C.....A.T.....G.....T..G.....A...C..C.G..A.....T....AC.....C....AT....C..
y11221 A.....T..T.....CCC...A.....G.....C..C.....C...C....A.....C.....A.T.....T.G...T.....TT.G..A..A...C..C....A.....C..T...AC.....C....A.....C..
8q243 .....T..T.....CCC.T.A...T.G.....C..C.....C...C....A.....C.....A.T.....C.....T..G.....A...C..C.G..A..T...A.T....AC.....C....A.....C..

7841 8000
HERV-KT47D TAAATTTCCACCATTGTGTGTAGGATTTGCACCTGGTTGTTTGGCATTCTCTAATCAAAATTGGATGTGGACTGTACCGGCCTCCAGCAATGATTCTTATCAGGTGCATAATGTCTTCTGTAGTAATTCTTTTCAGGTTCTGACCGTTAACATAAATTCA
10p151 .....
19p1311 .....A.....C.AC..A.T..GAAGT.G.....C.....A..T.AT...T.C..G.....C.....T...TT.CA.....A..A.....G...A...C...C.....AC....C..T..C..ACGT.CCC..
17q2131 .....A.....C.A..CACT..GA.GT.G.....C.....A.TT.AT...T.C..G.....A.....T...TT.CA.....A..A.....G...A....C.....AC....----.....AC.T...C..
16p1311 .....A.....A...A.T..GAAGT.G.....A..T.AT...T.A..G.....T...TT.CA.....A..AT.....G...A....C...C.....AC....C..T..C..ACGT.CCC..
4q131 .....AC.....C.....A.--------.....C.....A..T.AT...T.C..G.....C.....AT...TA.CA.....A..A.....G...A....C.....AC....C..T.....AC.T..GC.T
y11221 .....A..T....C.A..CA.T..GG.GTTG.....C.....A..T.AT...C.C..G.....CA....AT...TT.CA.....T.A..A.....G...A.G.CC.....AC....C..T..C..AC.TGC.C.T
8q243 .....A.....C.A..CA.T..GA.GT.TT....C.....A..T.AT...T.C..G.....C.....T...TT.CA.....GA..AT.....G...AA..CC.....A.....C..T..C..AC.T...C..

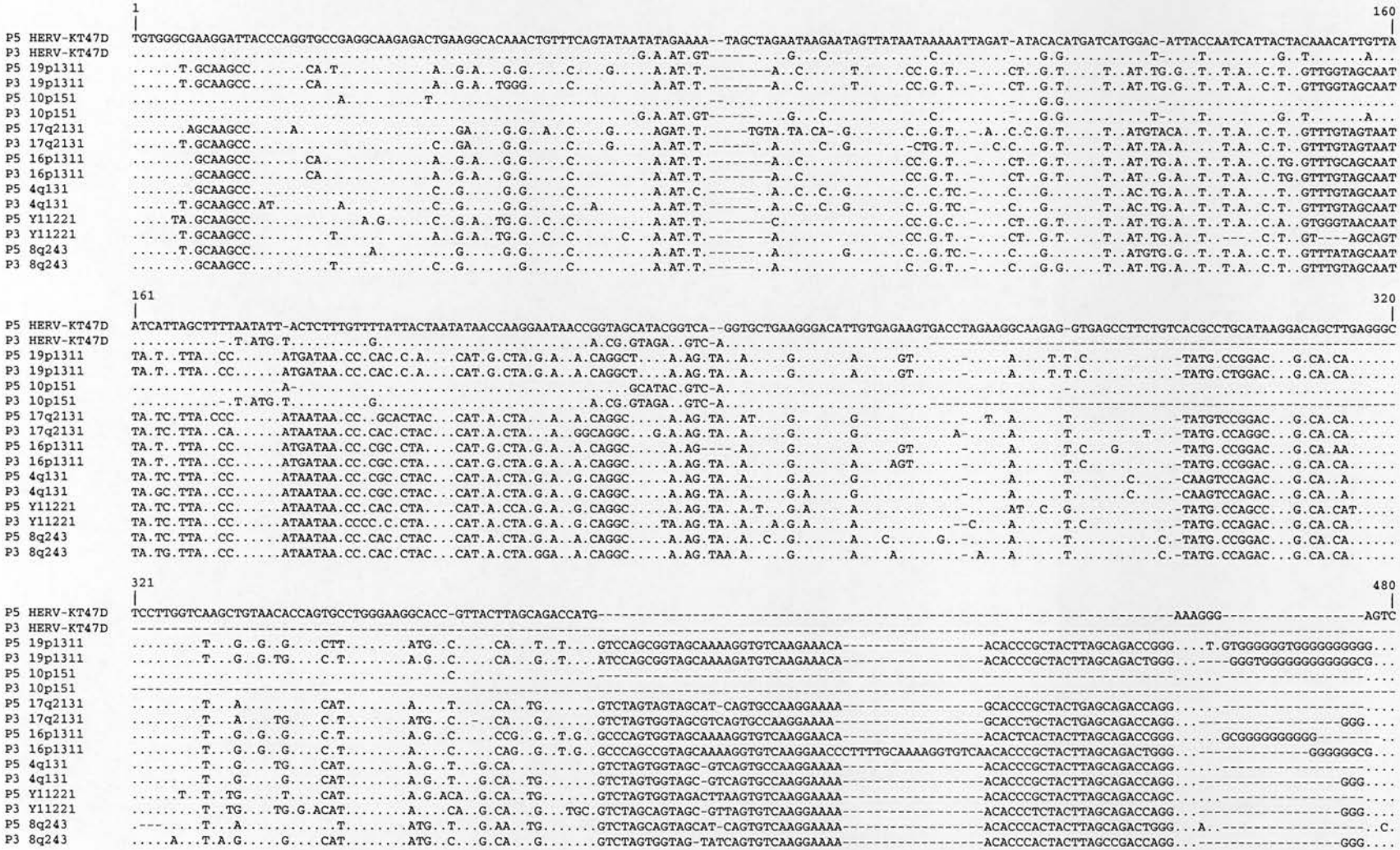
8001 8160
HERV-KT47D TTTGAAGAACAGAGAATTCCTGTCACAGTAAAGCATAATAAAACACAAGGATTGCCAGACTGTTTAAAAGACCTTATAAAGGGACCT--AATAATTCAAAACATTCTATGGAGTGATTGCAATGCCCCAAAAGTAGTAGTGCTAAGAAGTCCGATCACAA
10p151 .....--.....G.....G.....
19p1311 CA...G.C.TG.....C....TACC..AAC.....A.....C.....CC...GA.A.C..C.....TTT...G.GACTTCA...T.....GA.....T.....T..C..G.CT..T.....C.A..C..TAGC...GG
17q2131 CAA..G...TG.T.GG...TG.....TACT..CAGC.....A.....C.....CCC...G..ATC..C.....TTTC.GG.GA.TTCA...T.....A.A....T.....??????????????????????????????????
16p1311 CA...G.C.TG....G.....C....TACC..AAC.....A.....C....G.....CC...GA.A.C..C...T..G...TTT...G.GACTTCA...T.....GA.A....T.....T..C..G.CT..T..A..CCA..C..TAGC..TGG

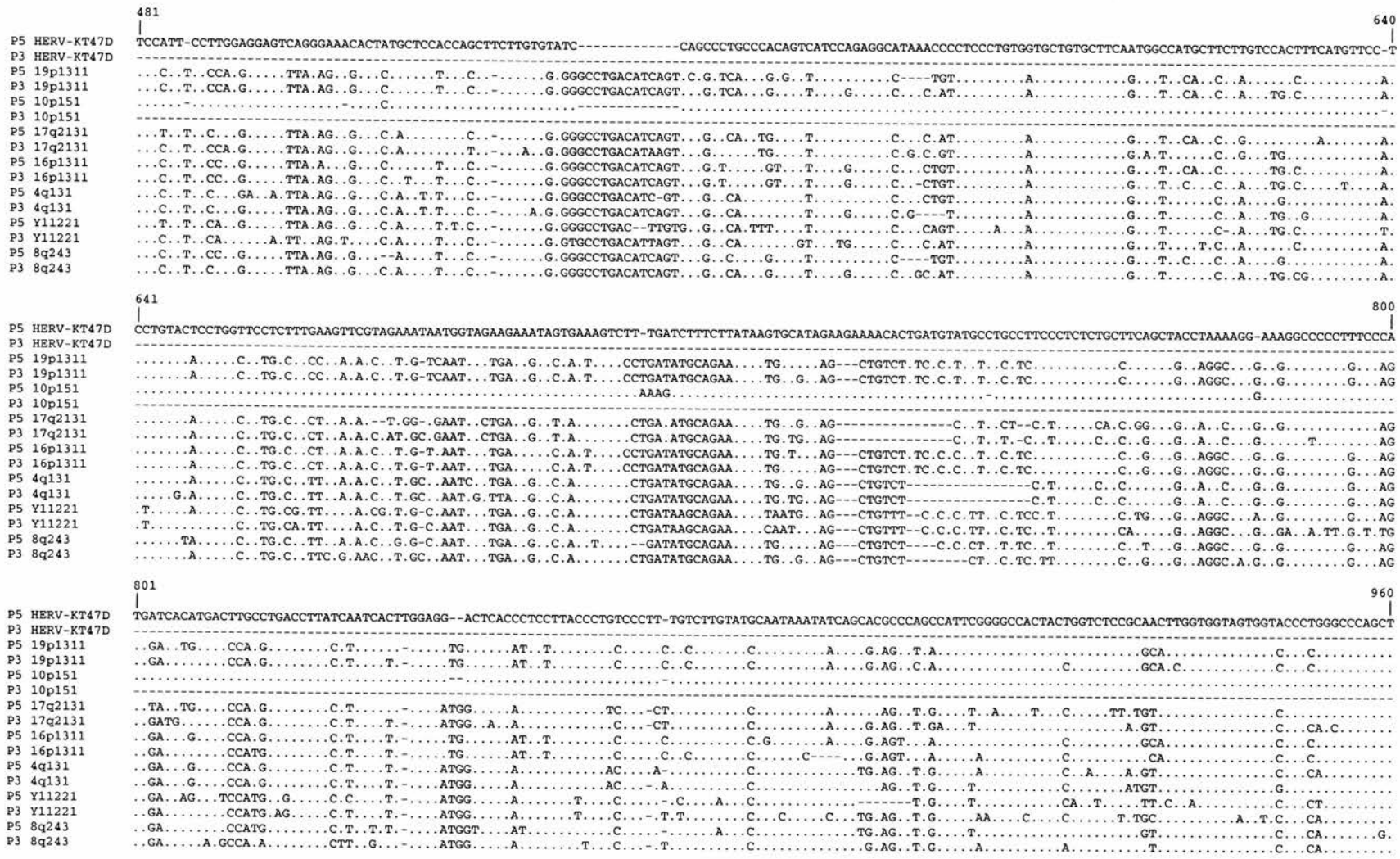

4q131
y11221
8q243
8161
8320
HERV-KT47D
10p151
19p1311
17q2131
16p1311
4q131
y11221
8q243
8321
8480
8481
8640
8641
8800
8801
8960
8961
9120

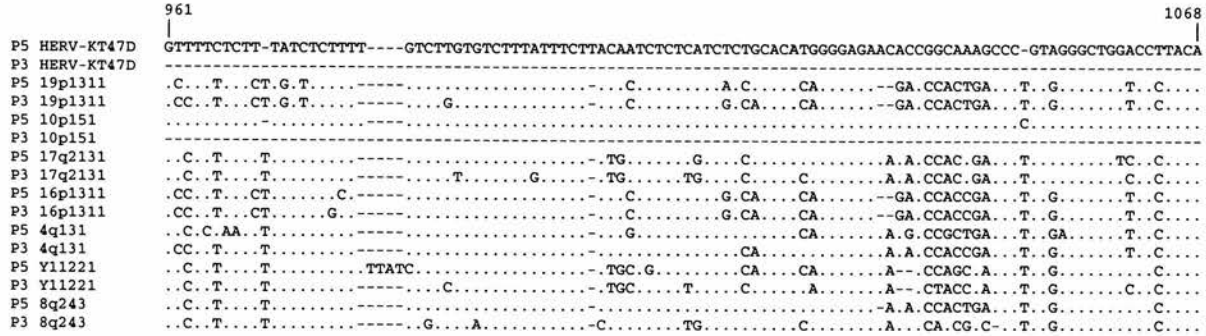



HERV-KT47D
10p151
19p1311
17q2131
16p1311
4q131
y11221
8q243
10081
10240
10241
10400
10401
10530

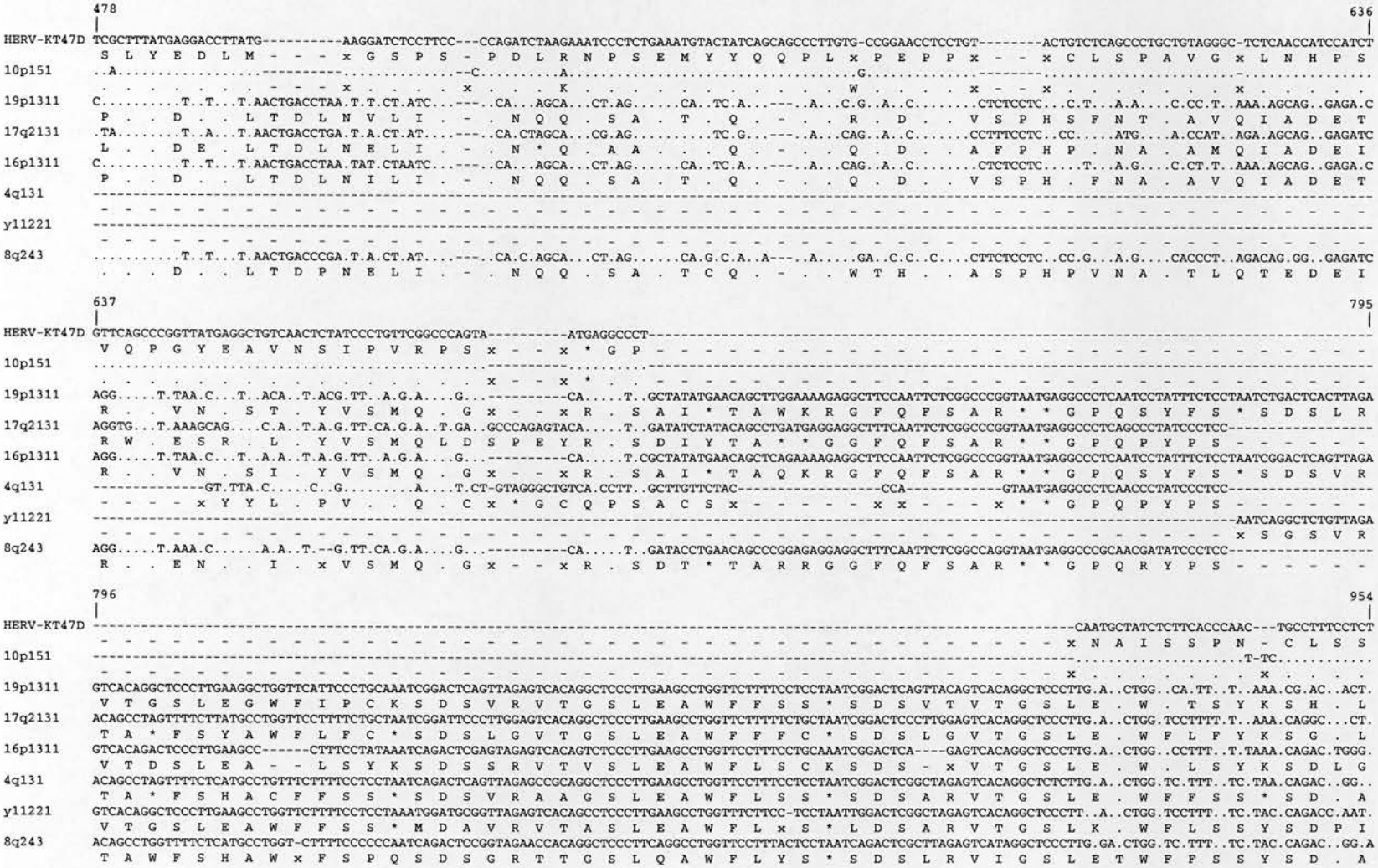
A.8 HERV-K(HML-4) Alignment of the respective LTRs of individual proviruses



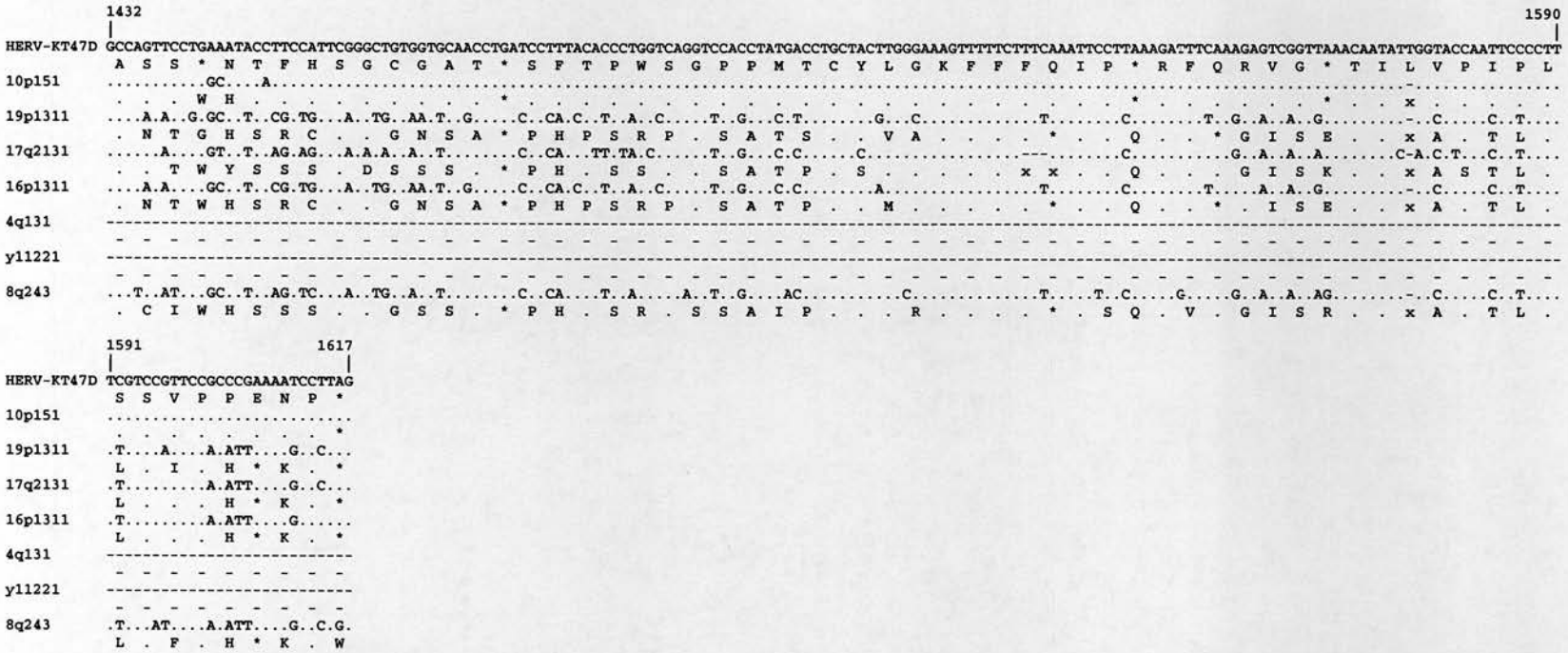

481
640
P5 HERV-KT47D
P3 HERV-KT47D
P5 19p1311
P3 19p1311
P5 10p151
P3 10p151
P5 17q2131
P3 17q2131
P5 16p1311
P3 16p1311
P5 4q131
P3 4q131
P5 Y11221
P3 Y11221
P5 8q243
P3 8q243
641
800
P5 HERV-KT47D
P3 HERV-KT47D
P5 19p1311
P3 19p1311
P5 10p151
P3 10p151
P5 17q2131
P3 17q2131
P5 16p1311
P3 16p1311
P5 4q131
P3 4q131
P5 Y11221
P3 Y11221
P5 8q243
P3 8q243
801
960
P5 HERV-KT47D
P3 HERV-KT47D
P5 19p1311
P3 19p1311
P5 10p151
P3 10p151
P5 17q2131
P3 17q2131
P5 16p1311
P3 16p1311
P5 4q131
P3 4q131
P5 Y11221
P3 Y11221
P5 8q243
P3 8q243


961
1068
P5 HERV-KT47D GTTTTCTCTT-TATCTCTTTT----GTCTTGTGTCTTTATTTCTTACAATCTCTCATCTCTGCACATGGGGAGAACACCGGCAAAGCCC-GTAGGGCTGGACCTTACA
P3 HERV-KT47D
P5 19p1311 .C...T...CT.G.T.....-----.....-...C.....A.C.....CA.....--GA.CCACTGA...T..G.....T..C....
P3 19p1311 .CC..T...CT.G.T.....-----....G.....-...C.....G.CA....CA.....--GA.CCACTGA...T..G.....T..C....
P5 10p151 .....-.....----.....C.....
P3 10p151
P5 17q2131 ..C..T....T.....-----.....-.TG.....G....C.....A.A.CCAC.GA...T.....TC..C....
P3 17q2131 ..C..T....T.....-----.....T.....G.....-.TG.....TG....C.....C.....A.A.CCAC.GA...T.....C..C....
P5 16p1311 .CC..T...CT.....C.-----.....-...C.....G.CA....CA.....--GA.CCACCGA...T..G.....T..C....
P3 16p1311 .CC..T...CT.....G..-----.....-...C.....G.CA....CA.....--GA.CCACCGA...T..G.....T..C....
P5 4q131 ..C.C.AA..T.....-----.....-...G.....CA.....A.G.CCGCTGA...T..GA.....T..C....
P3 4q131 .CC..T....T.....-----.....-.....CA.....A.A.CCACCGA...T..G.....T..C....
P5 Y11221 ..C..T....T.....TTATC.....-.TGC.G.....CA....CA.....A--.CCAGC.A...T..G.....C....
P3 Y11221 ..C..T....T.....-----....C.....-.TGC.....T.....C.....A.....A--.CTACC.A...T..G.....C..C....
P5 8q243 ..C..T....T.....-----.....-.....-A.A.CCACTGA...T..G.....C....
P3 8q243 ..C..T....T.....-----..G....A.....-C.....TG.....C.....A...CA.CG.C-..T..G.....C....

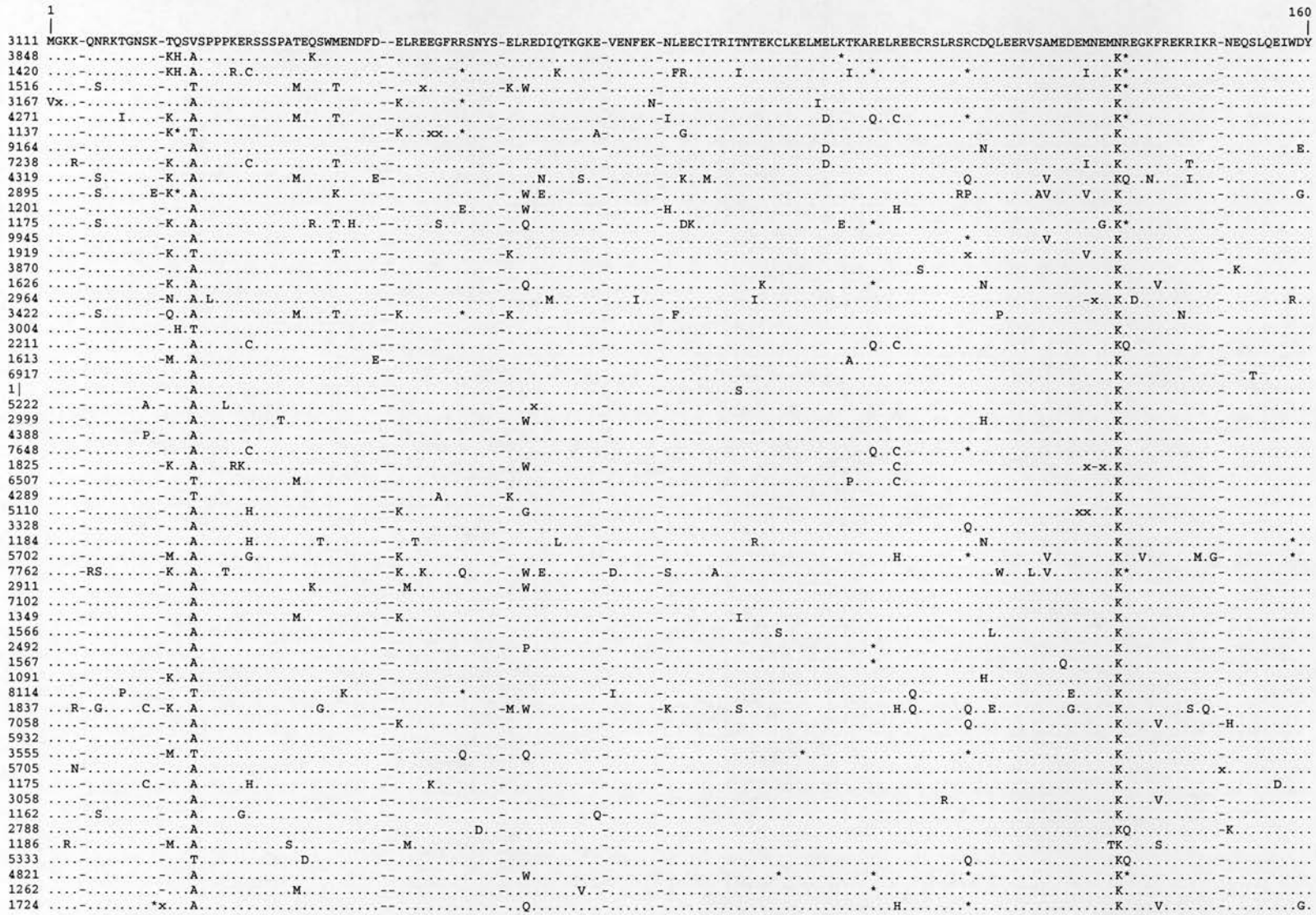
A.9 HERV-K(HML-4) Alignment of the Gag ORFs

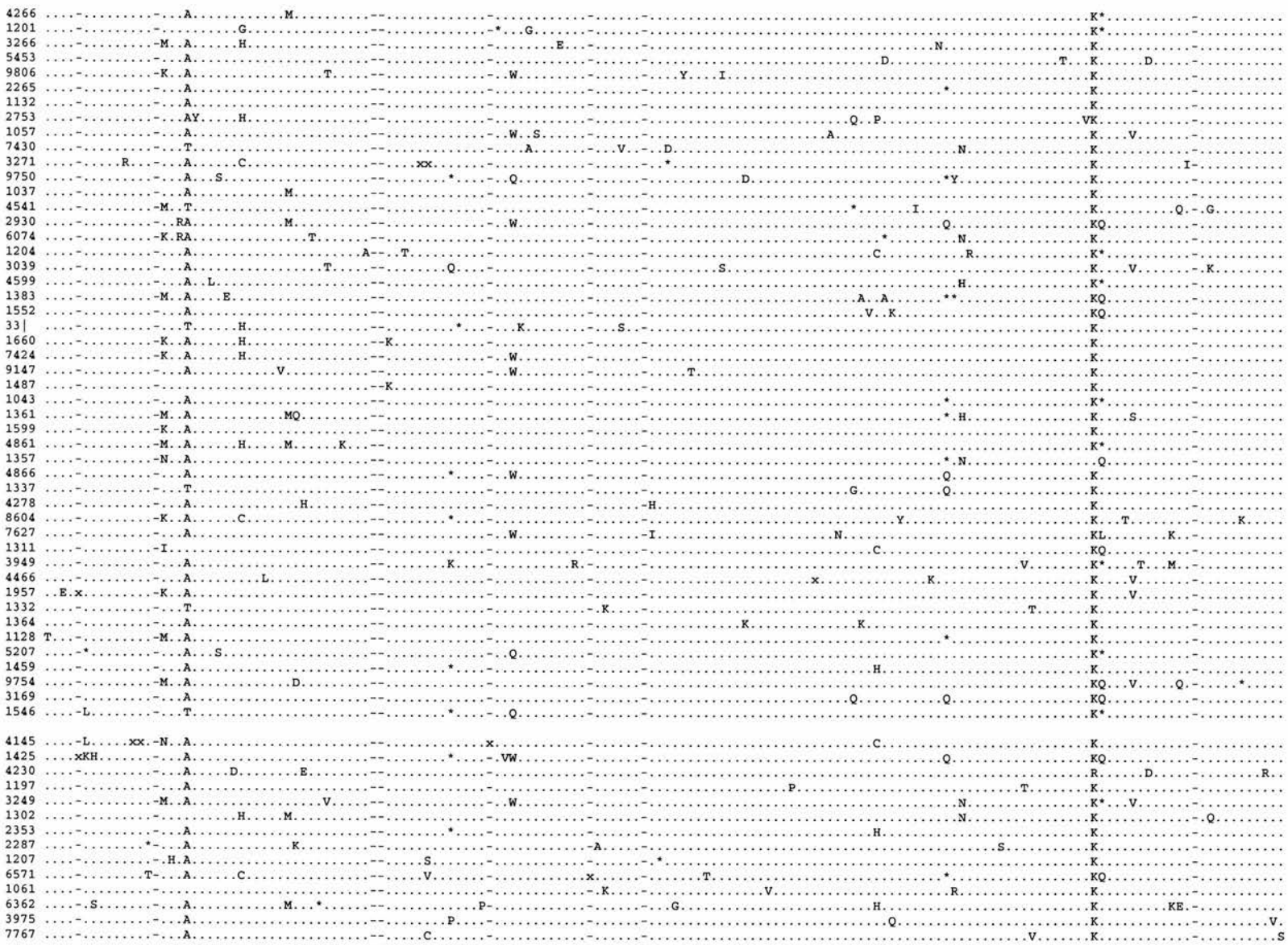

478
636
637
795
796
954
HERV-KT47D
10p151
19p1311
17q2131
16p1311
4q131
y11221
8q243

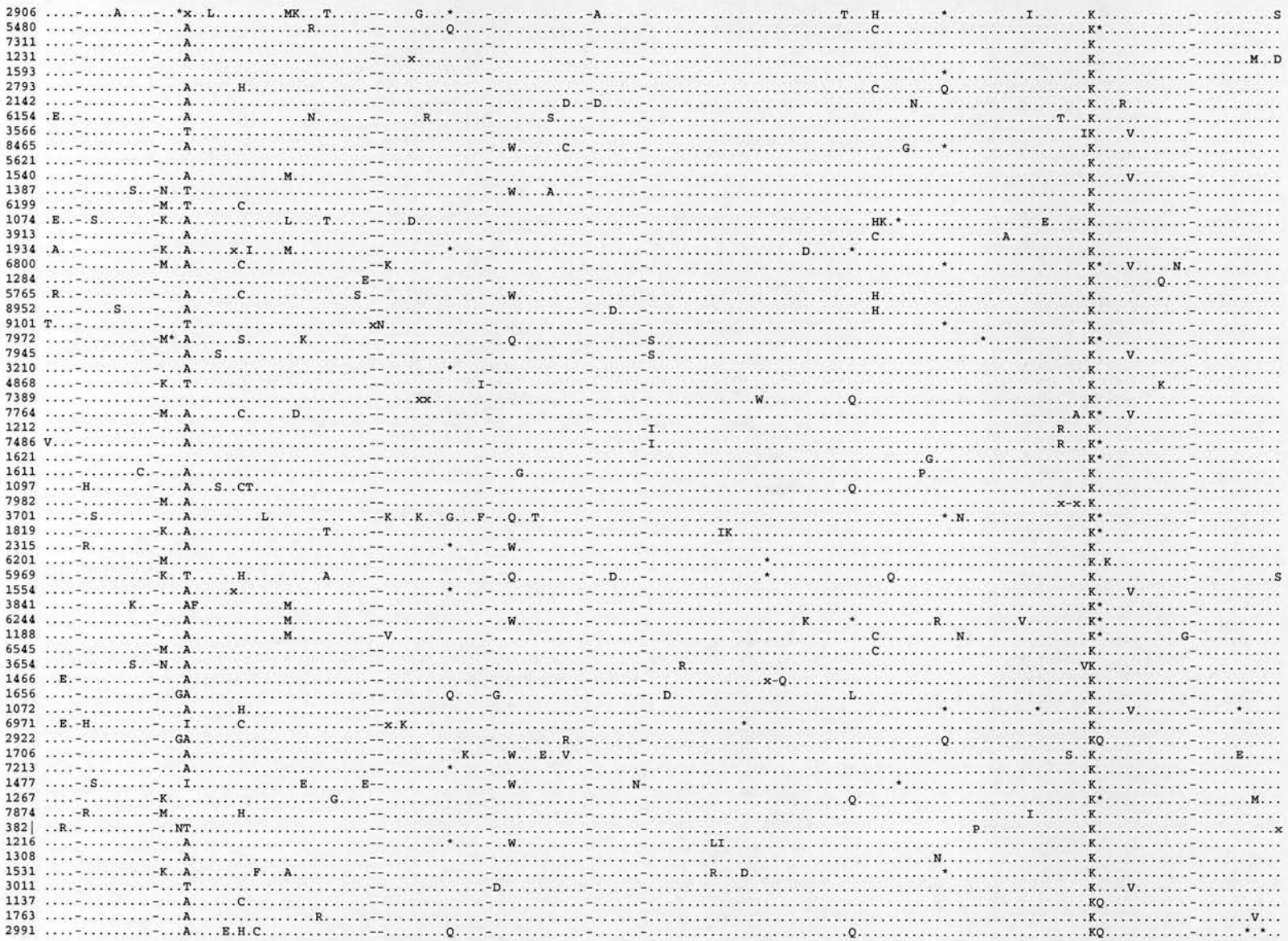

955
1113
HERV-KT47D
10p151
19p1311
17q2131
16p1311
4q131
y11221
8q243
1114
1272
1273
1431

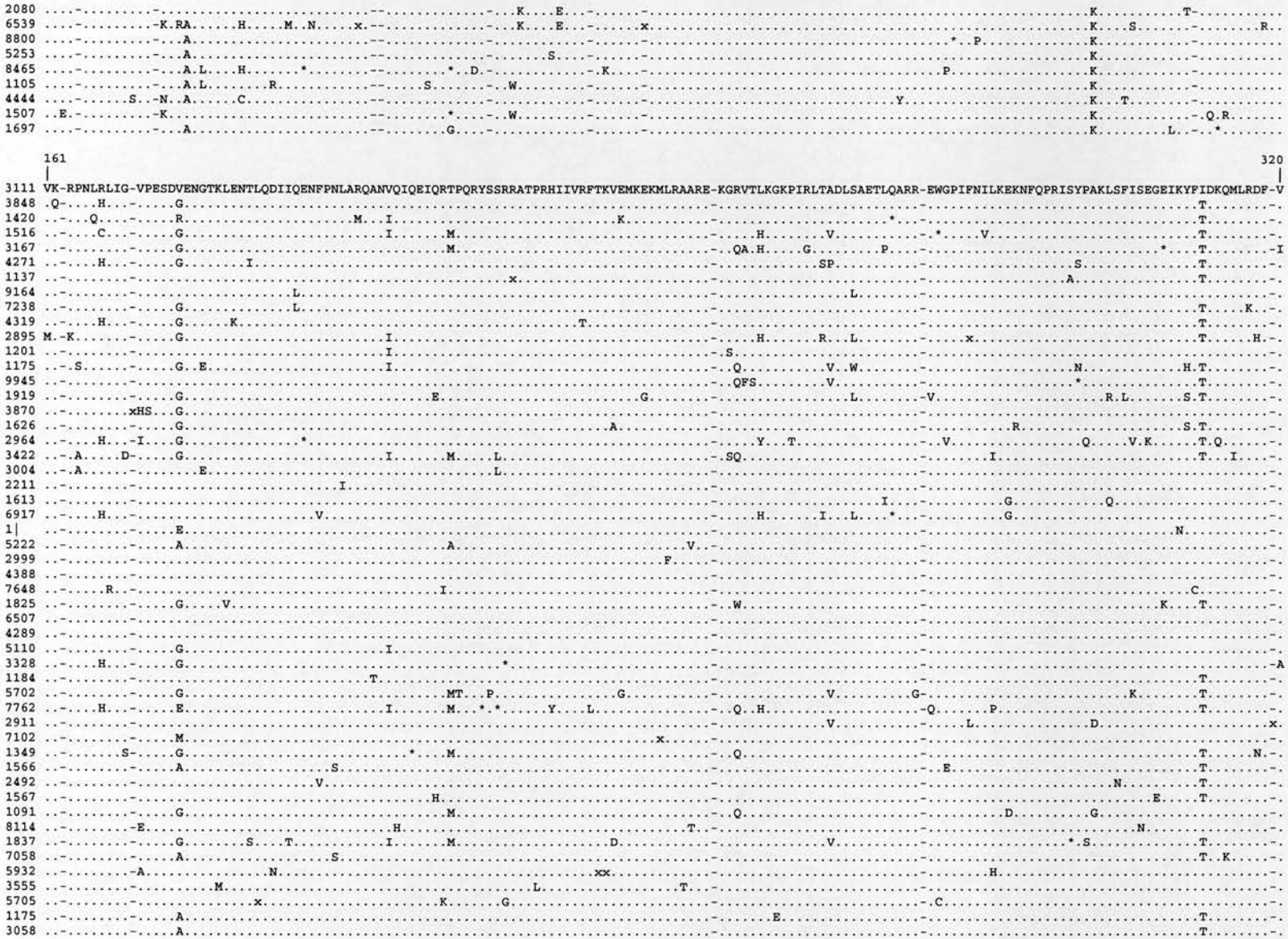

1432
1590
HERV-KT47D GCCAGTTCCTGAAATACCTTCCATTCGGGCTGTGGTGCAACCTGATCCTTTACACCCTGGTCAGGTCCACCTATGACCTGCTACTTGGGAAAGTTTTTCTTTCAAATTCCTTAAAGATTTCAAAGAGTCGGTTAAACAATATTGGTACCAATTCCCCTT
A S S * N T F H S G C G A T * S F T P W S G P P M T C Y L G K F F F Q I P * R F Q R V G * T I L V P I P L
10p151
19p1311
17q2131
16p1311
4q131
y11221
8q243
1591
1617
HERV-KT47D TCGTCCGTTCCGCCCGAAAATCCTTAG
S S V P P E N P *
10p151
19p1311
17q2131
16p1311
4q131
y11221
8q243

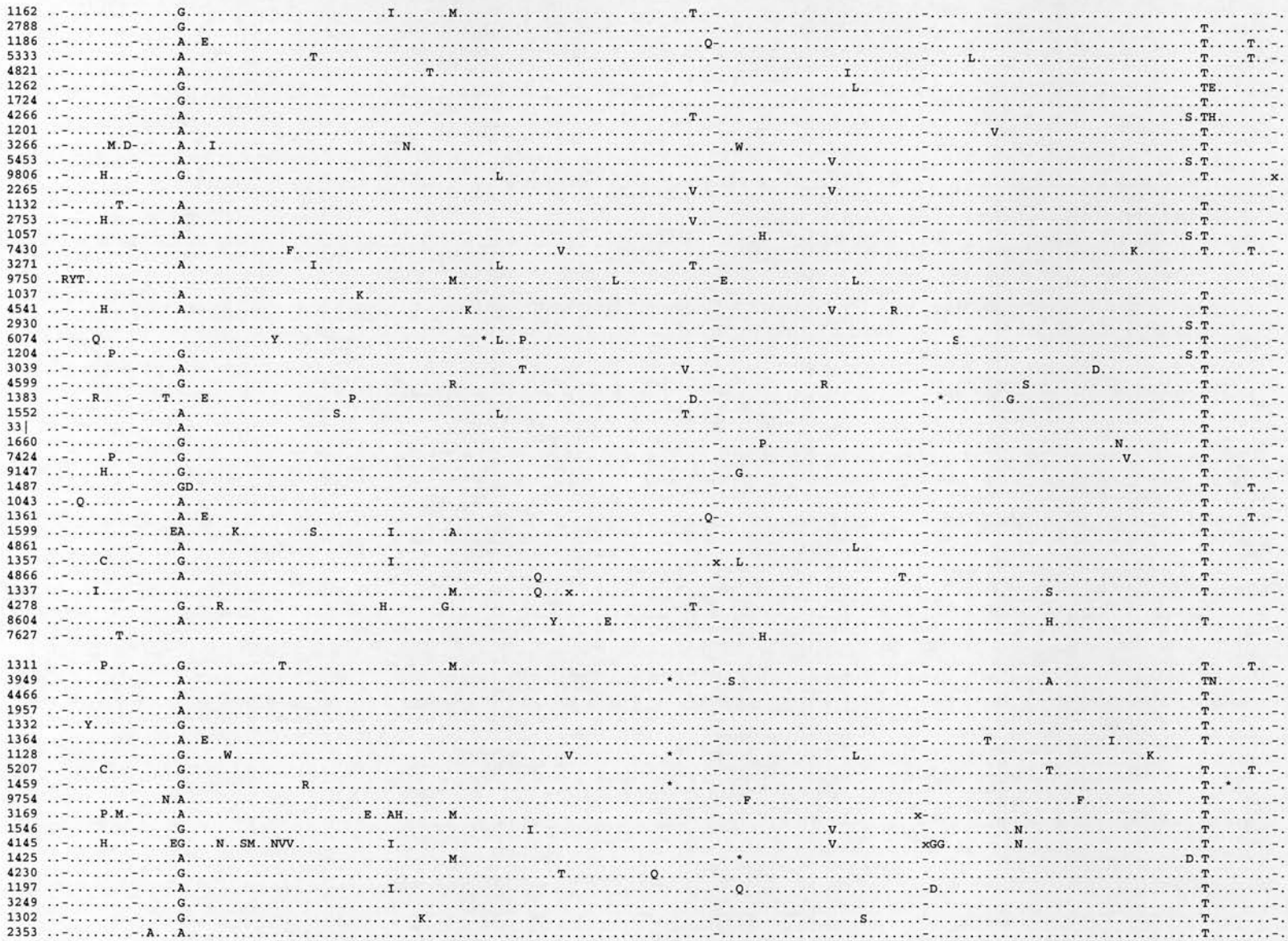
A.10 LINE elements alignment of ORFs 1 and 2

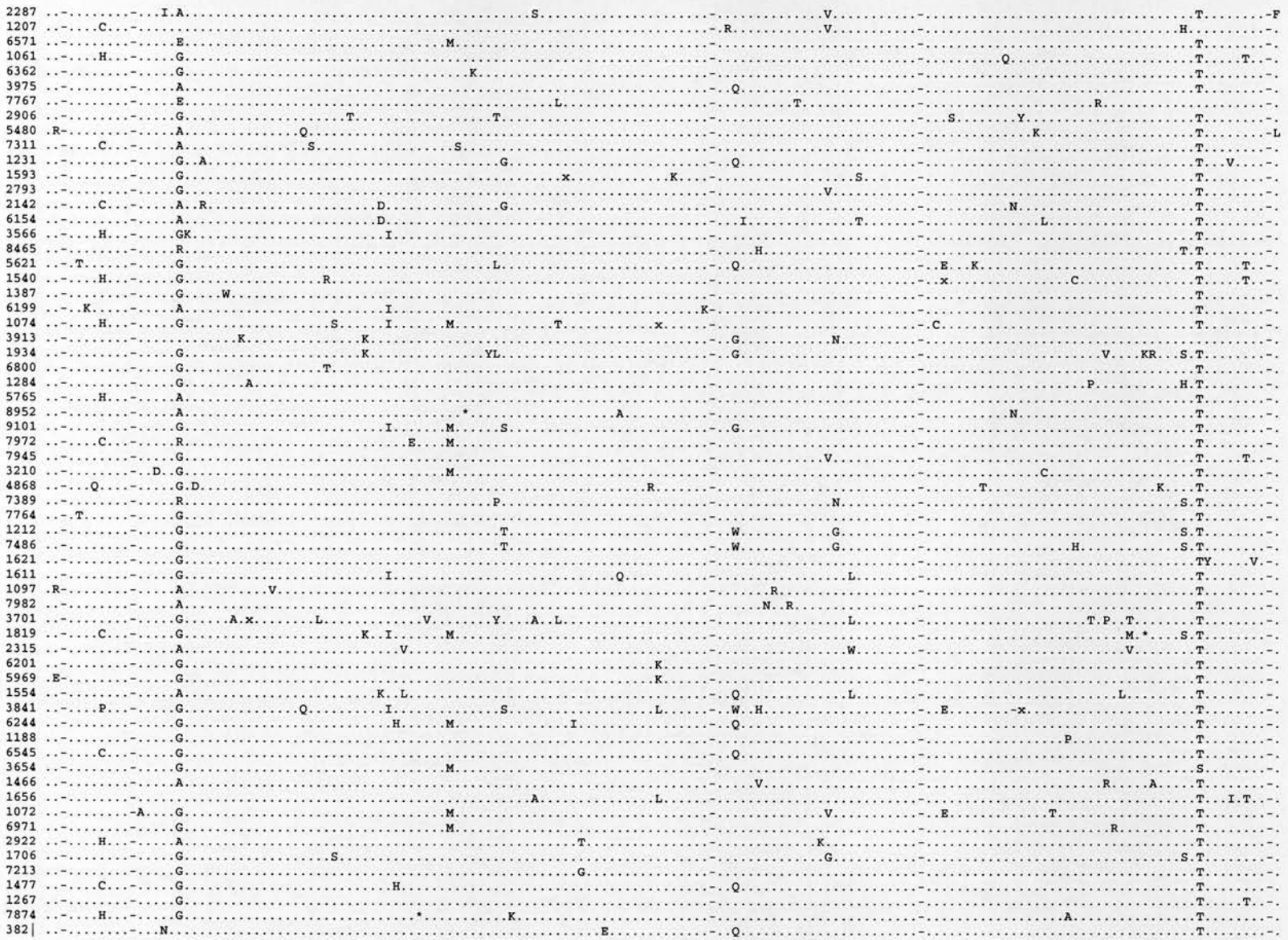


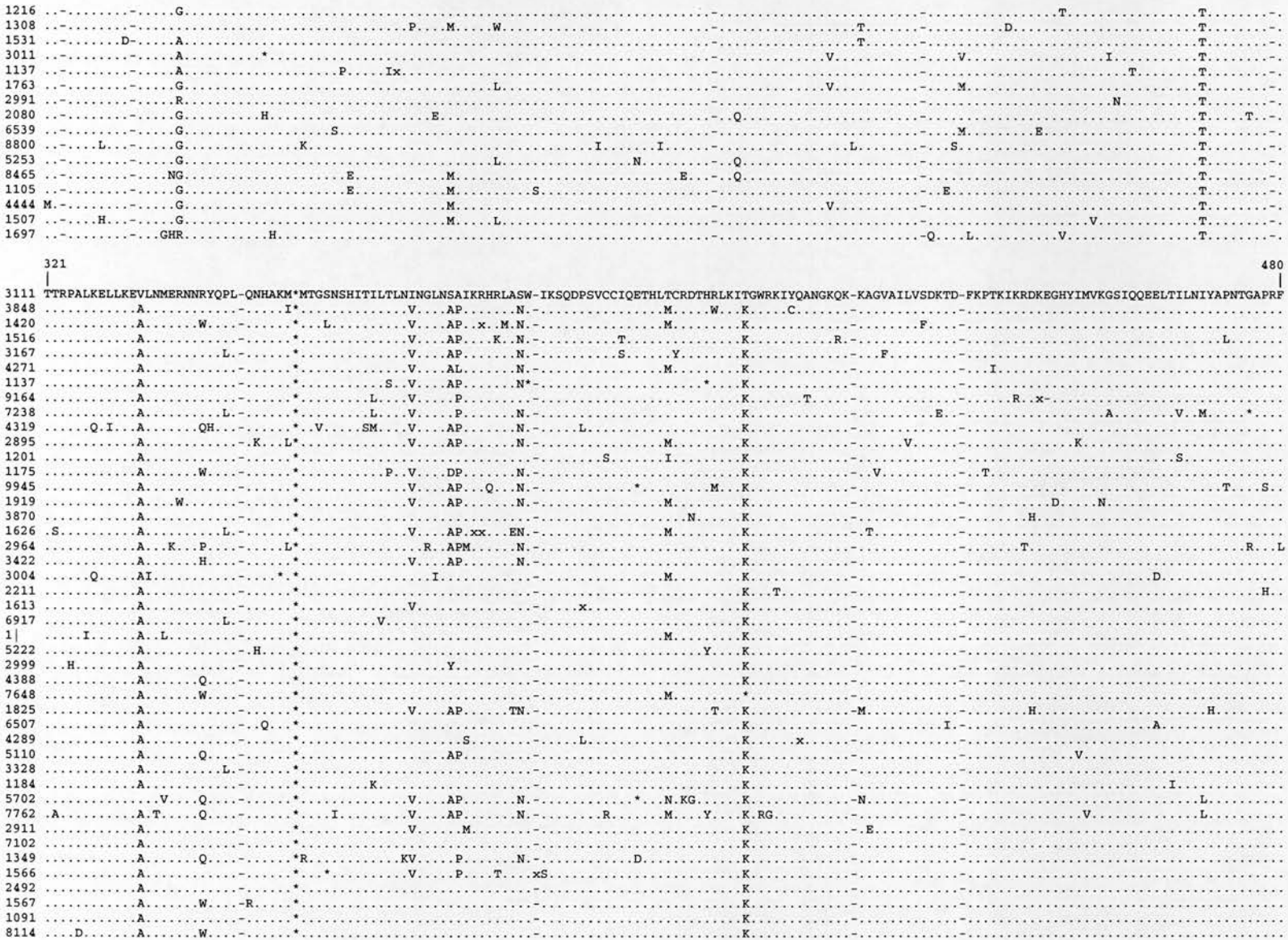


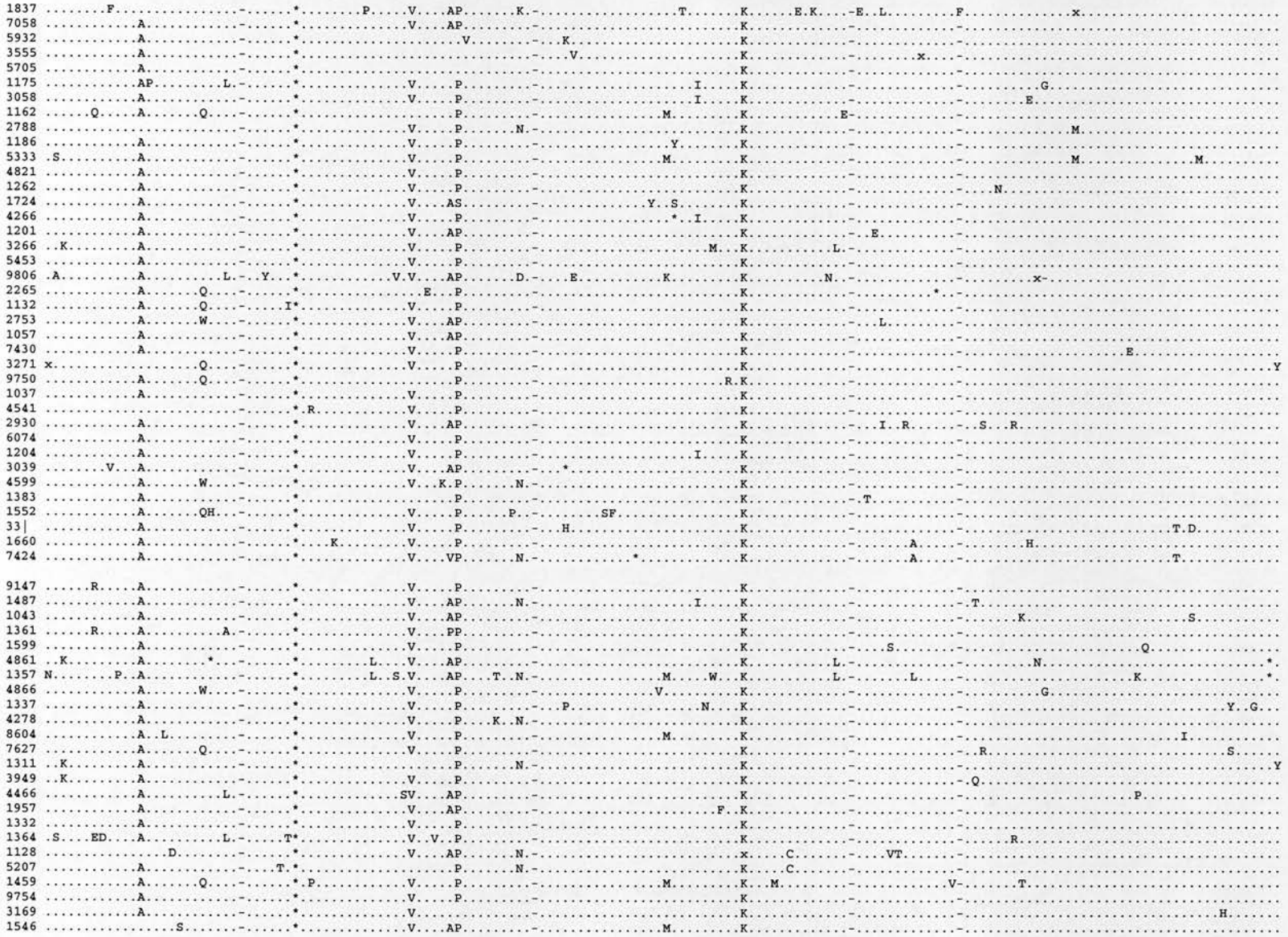


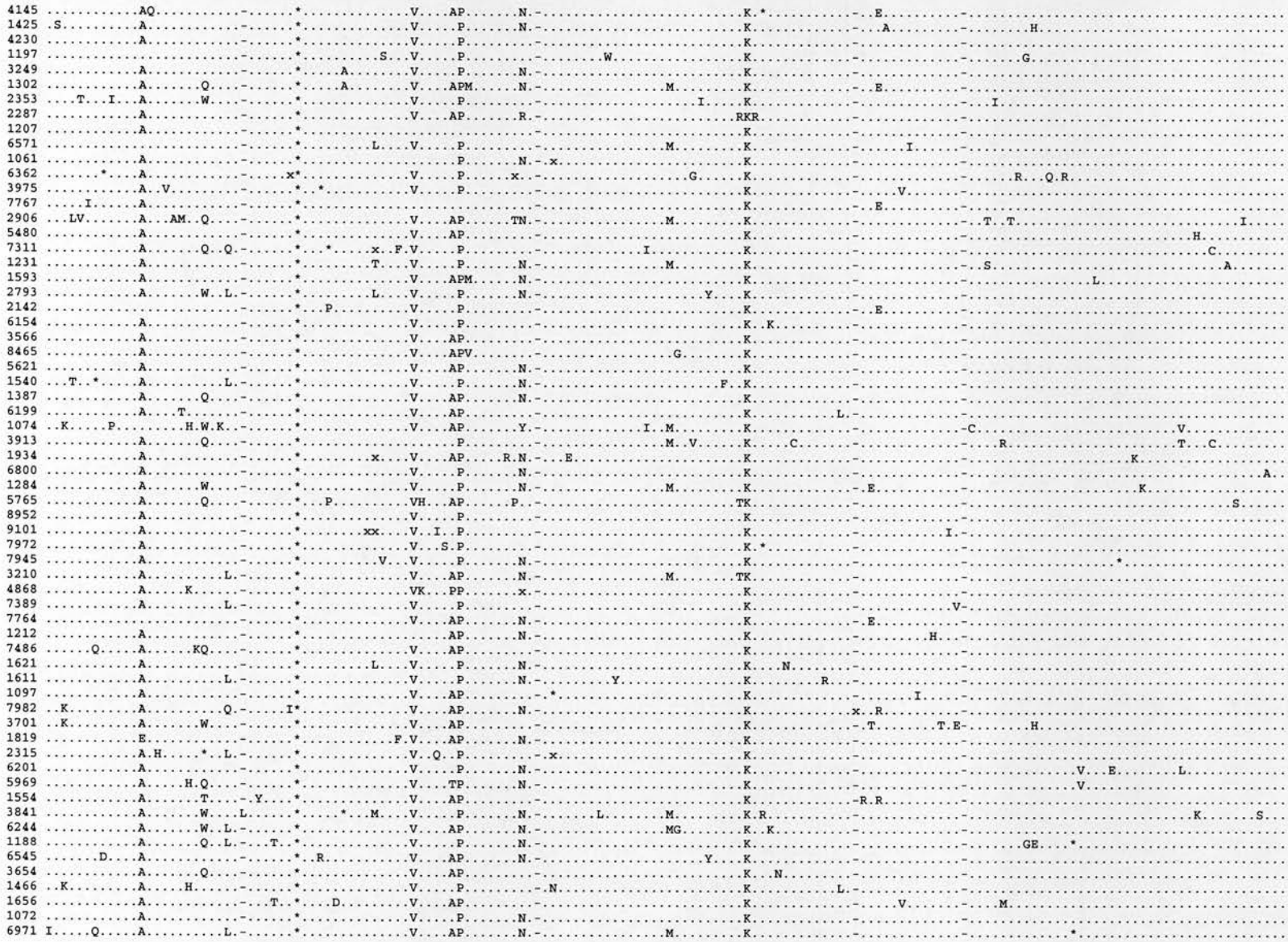


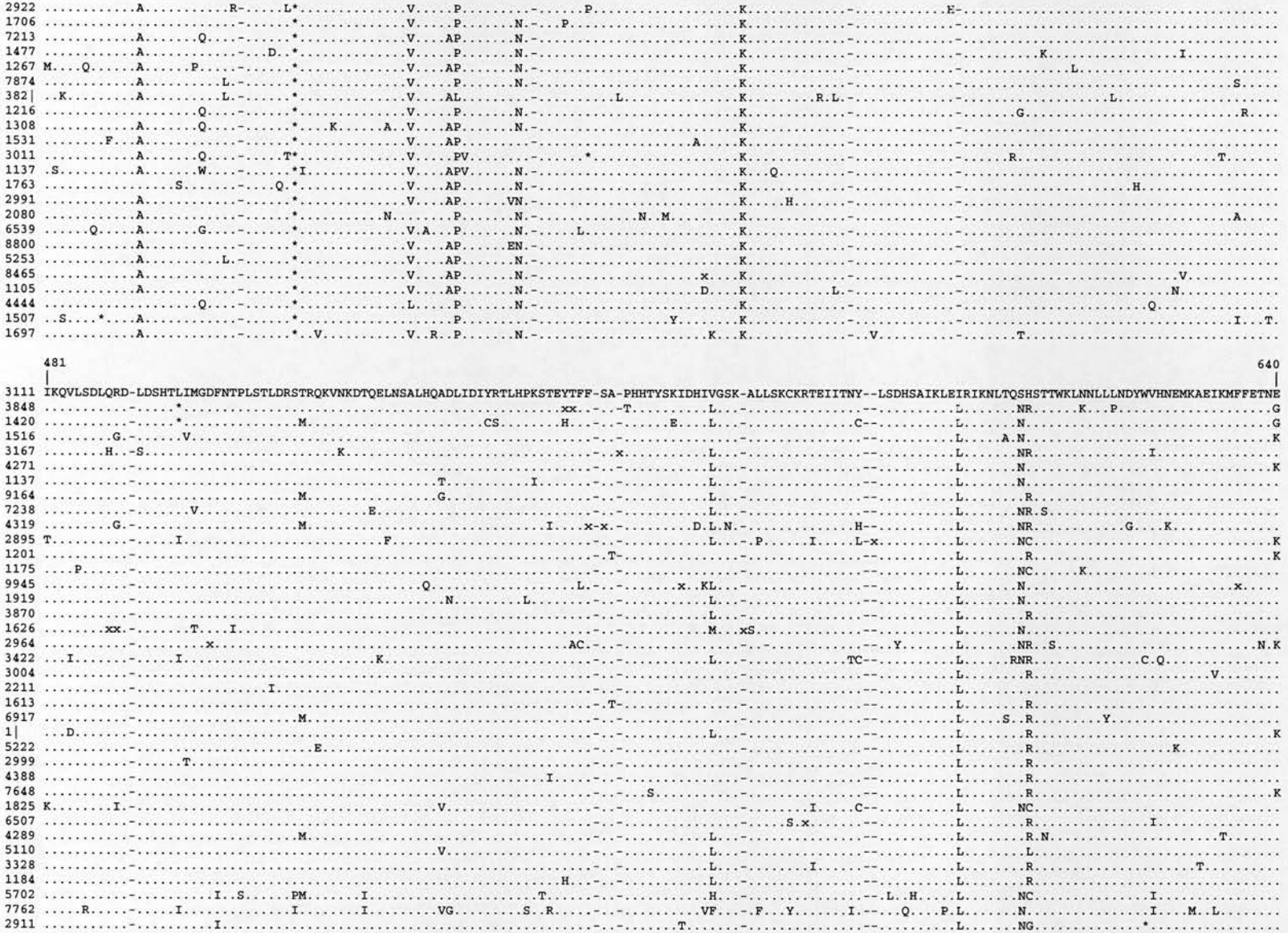


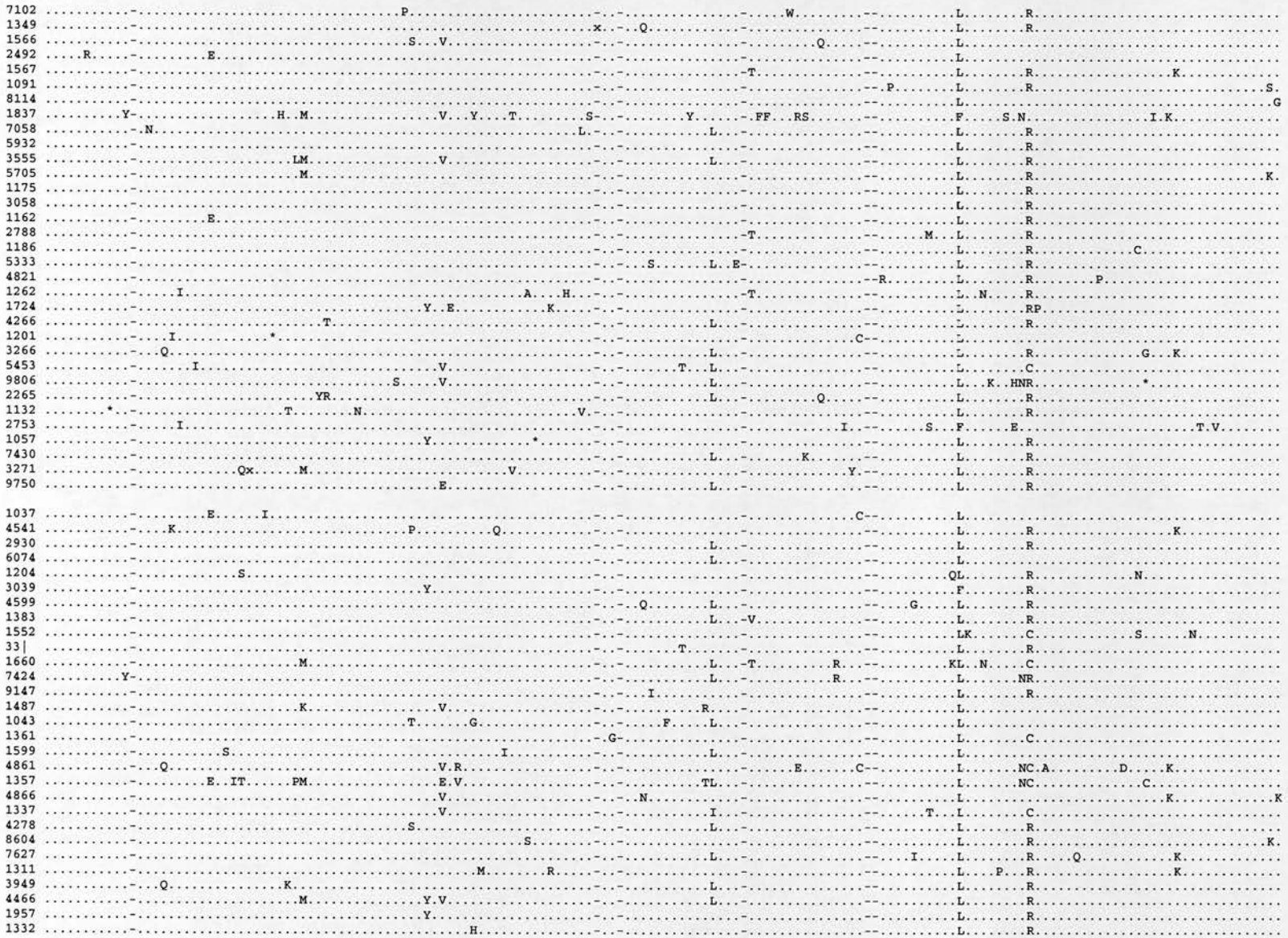


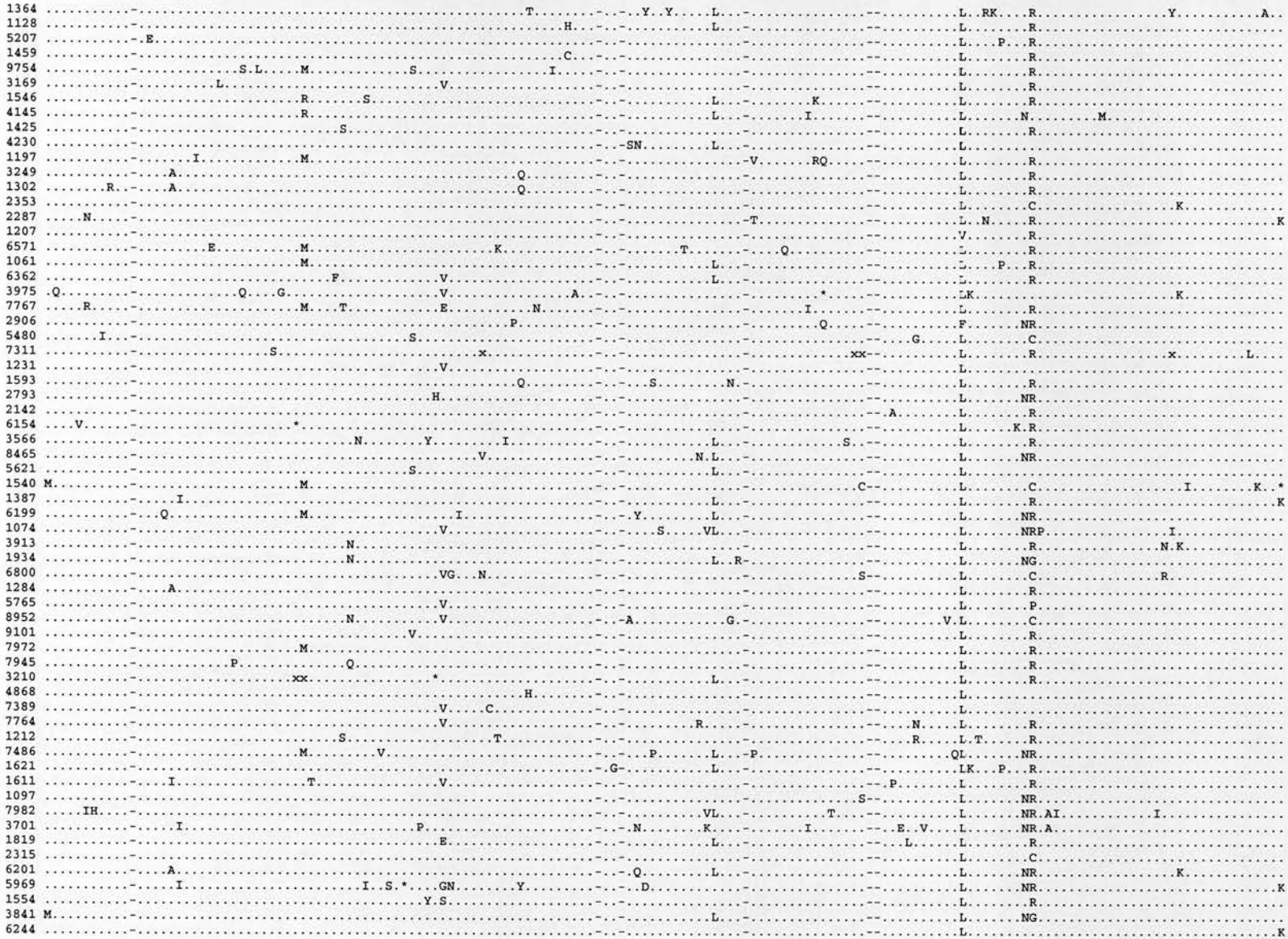


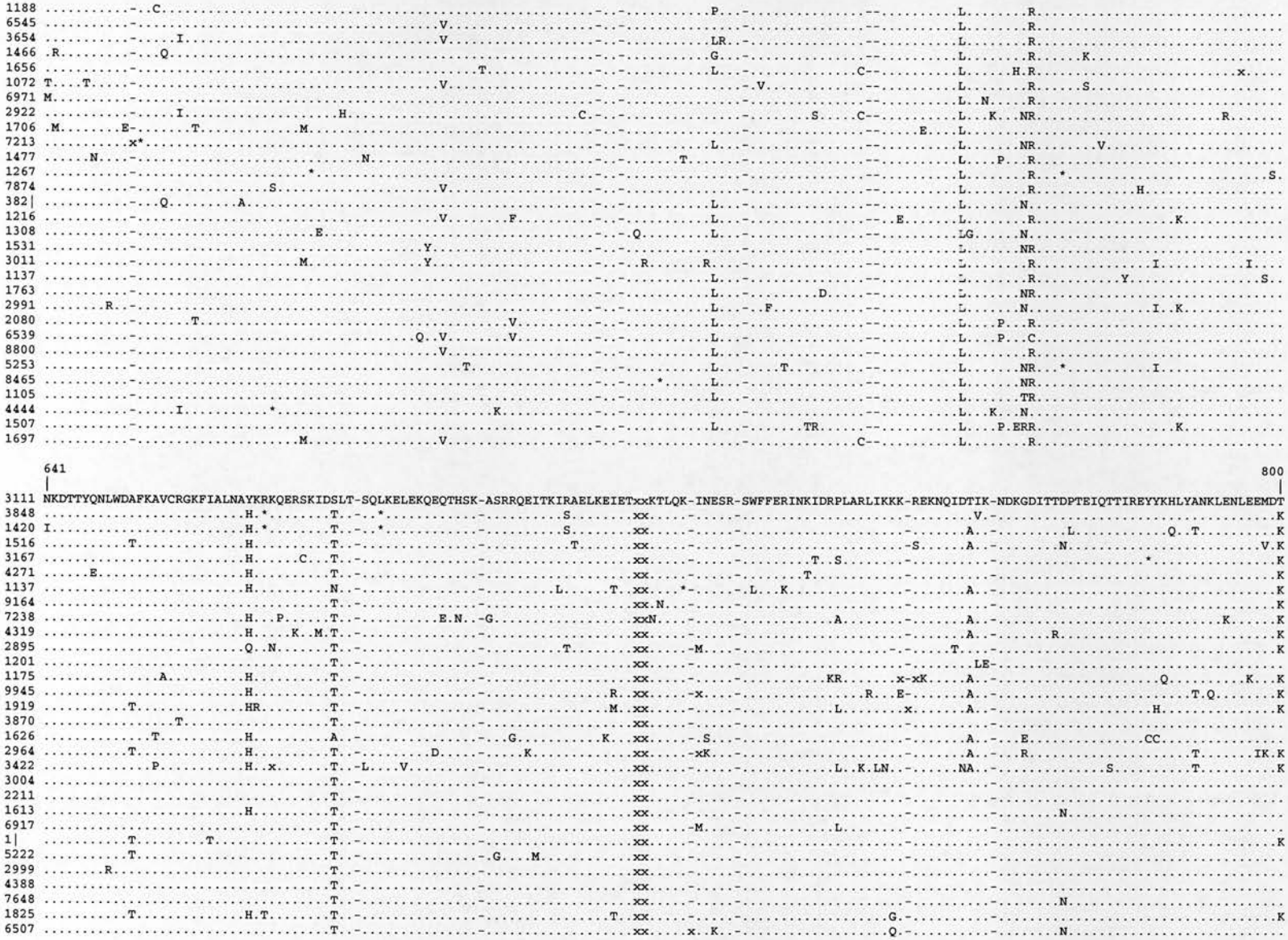


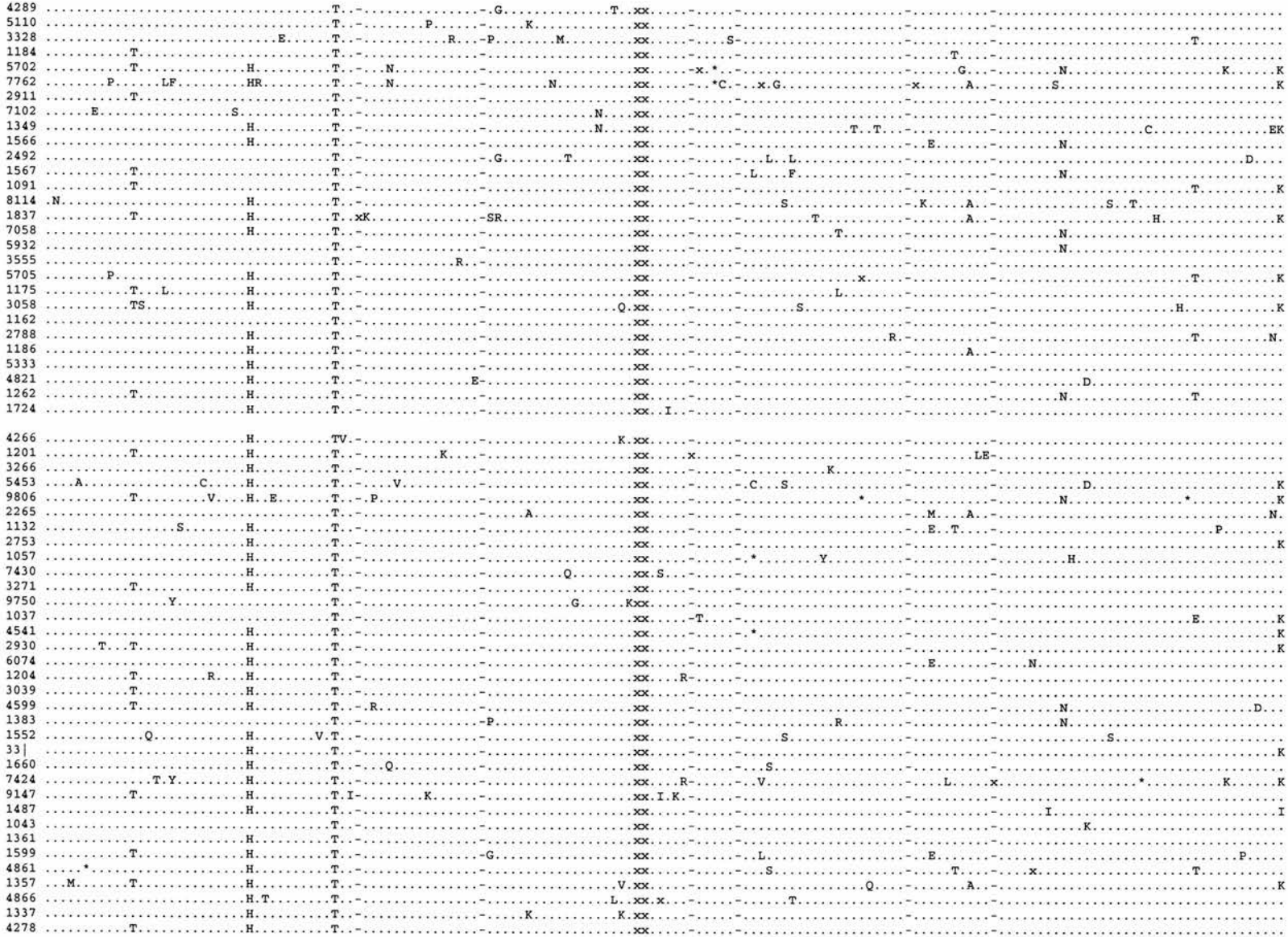


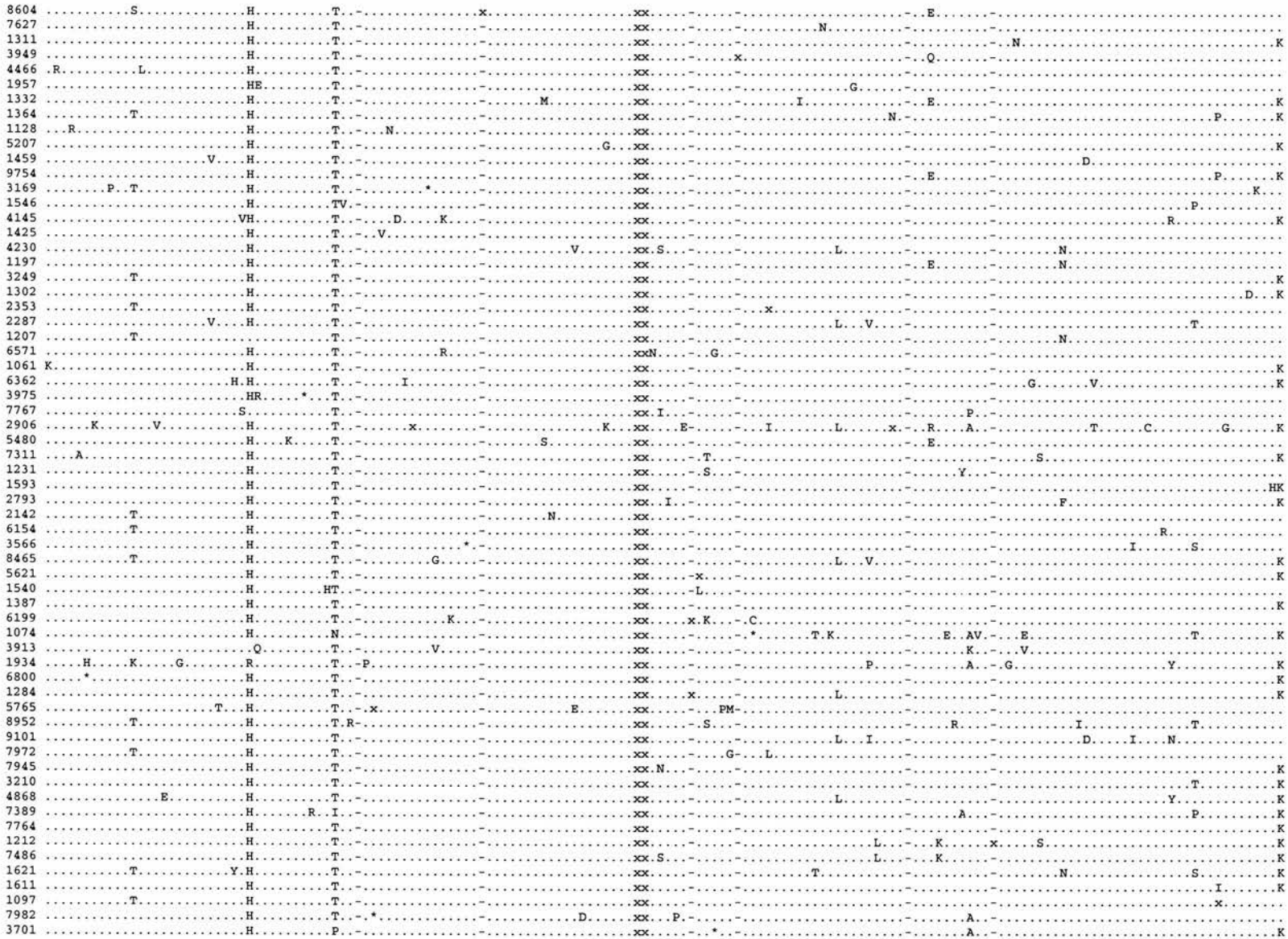

481
640
IKQVLSDLQRD-LDSHTLIMGDFNTPLSTLDRSTRQKVNKDTQELNSALHQADLIDIYRTLHPKSTEYTFF-SA-PHHTYSKIDHIVGSK-ALLSKCKRTEIITNY--LSDHSAIKLEIRIKNLTQSHSTTWKLNNLLLNDYWVHNEMKAEIKMFFETNE

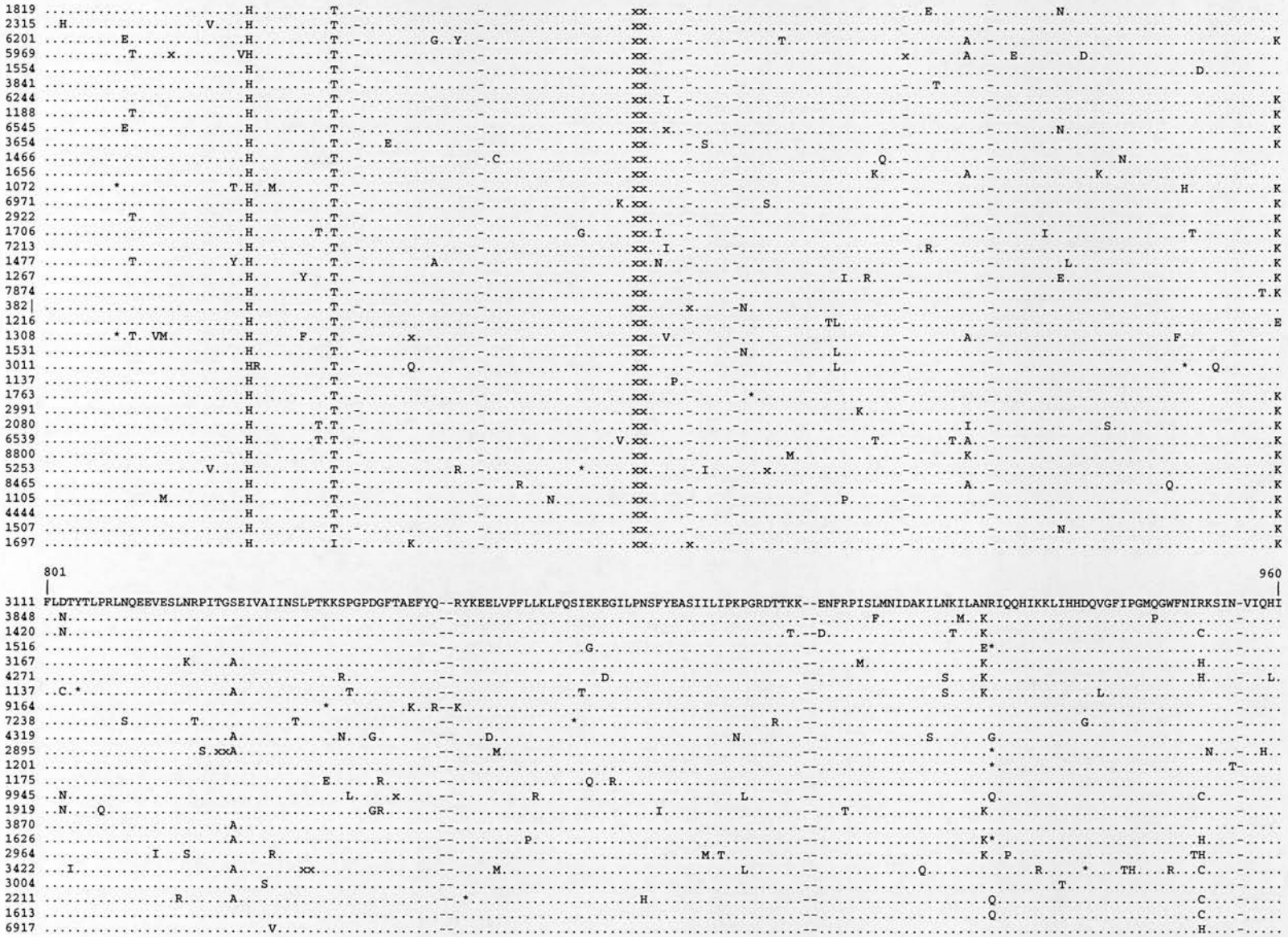


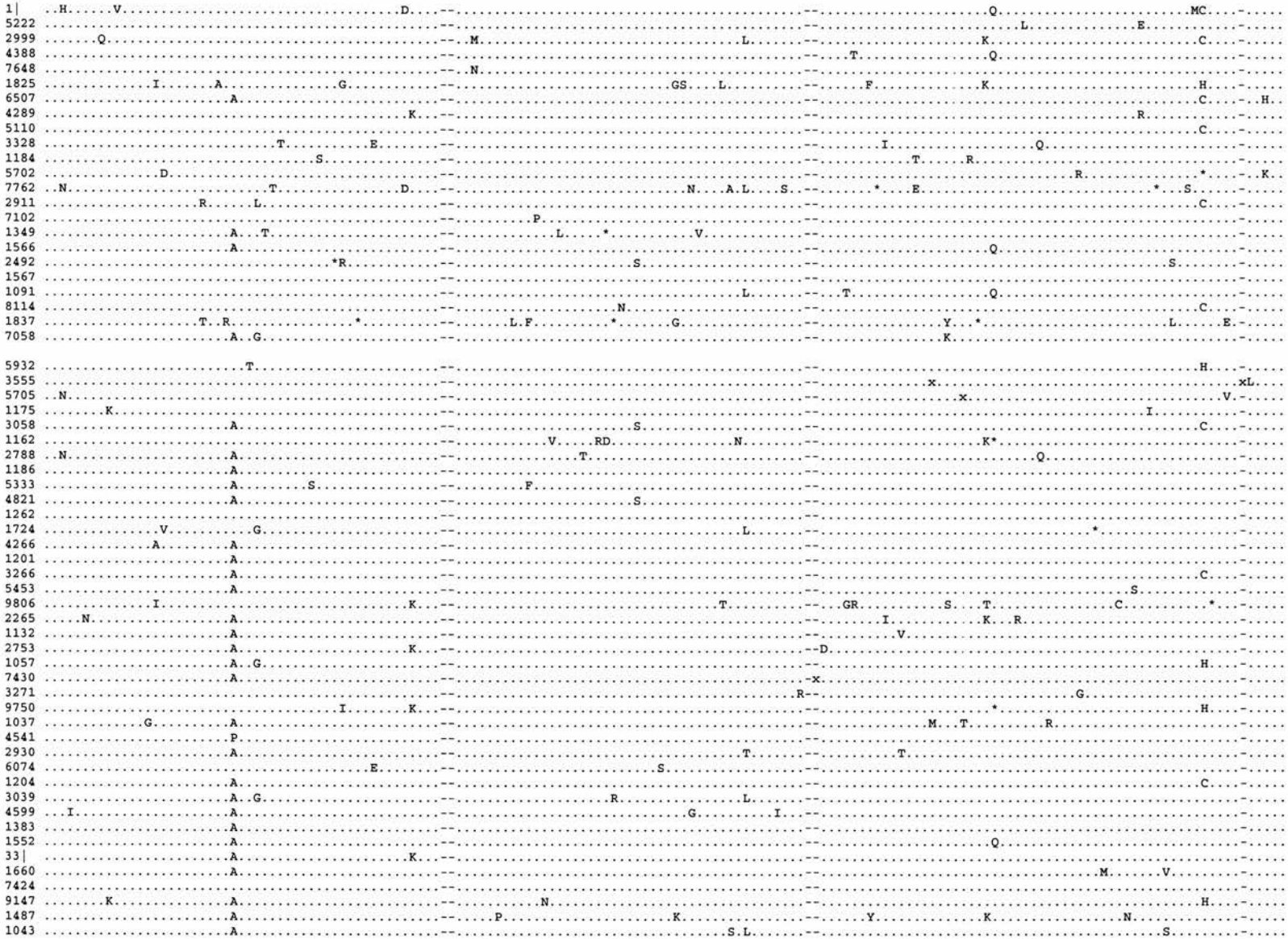


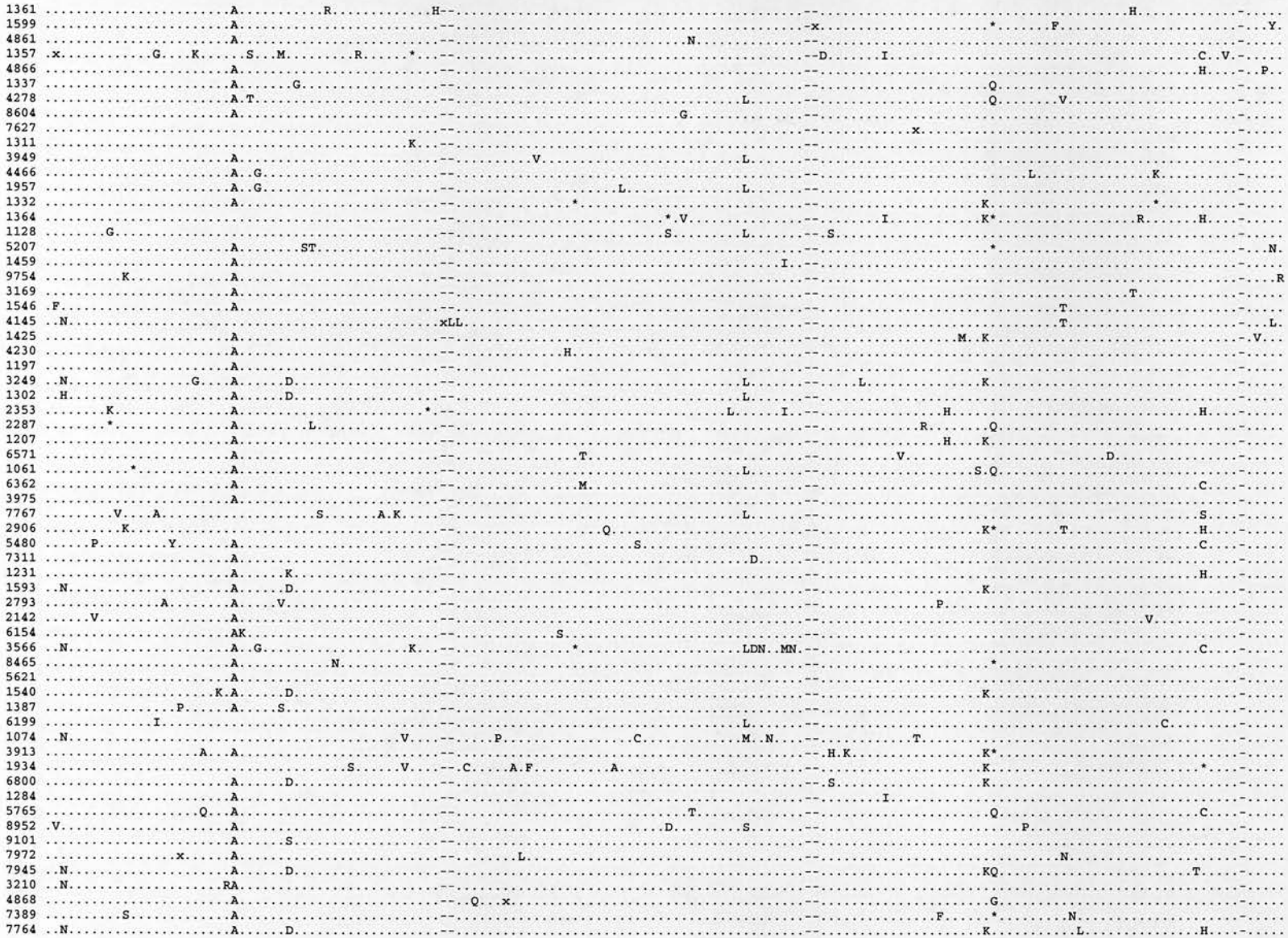

641
800
3111 NKDTTYQNLWDAFKAVCRGKFIALNAYKRKQERSKIDSLT-SQLKELEKQEQTHSK-ASRRQEITKIRAELKEIETxxKTLQK-INESR-SWFFERINKIDRPLARLIKKK-REKNQIDTIK-NDKGDITTDPTEIQTTIREYYKHLYANKLENLEEMDT

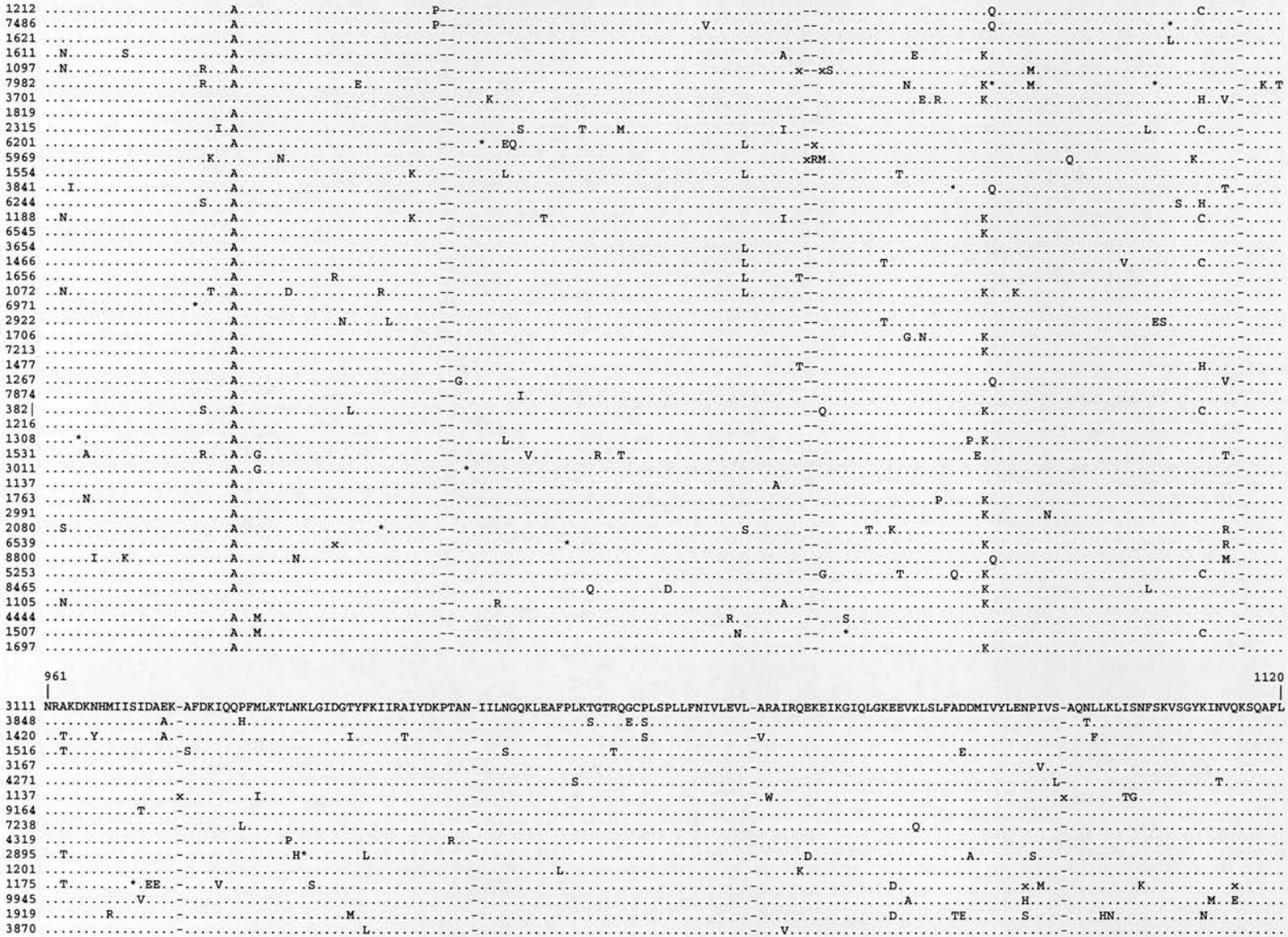


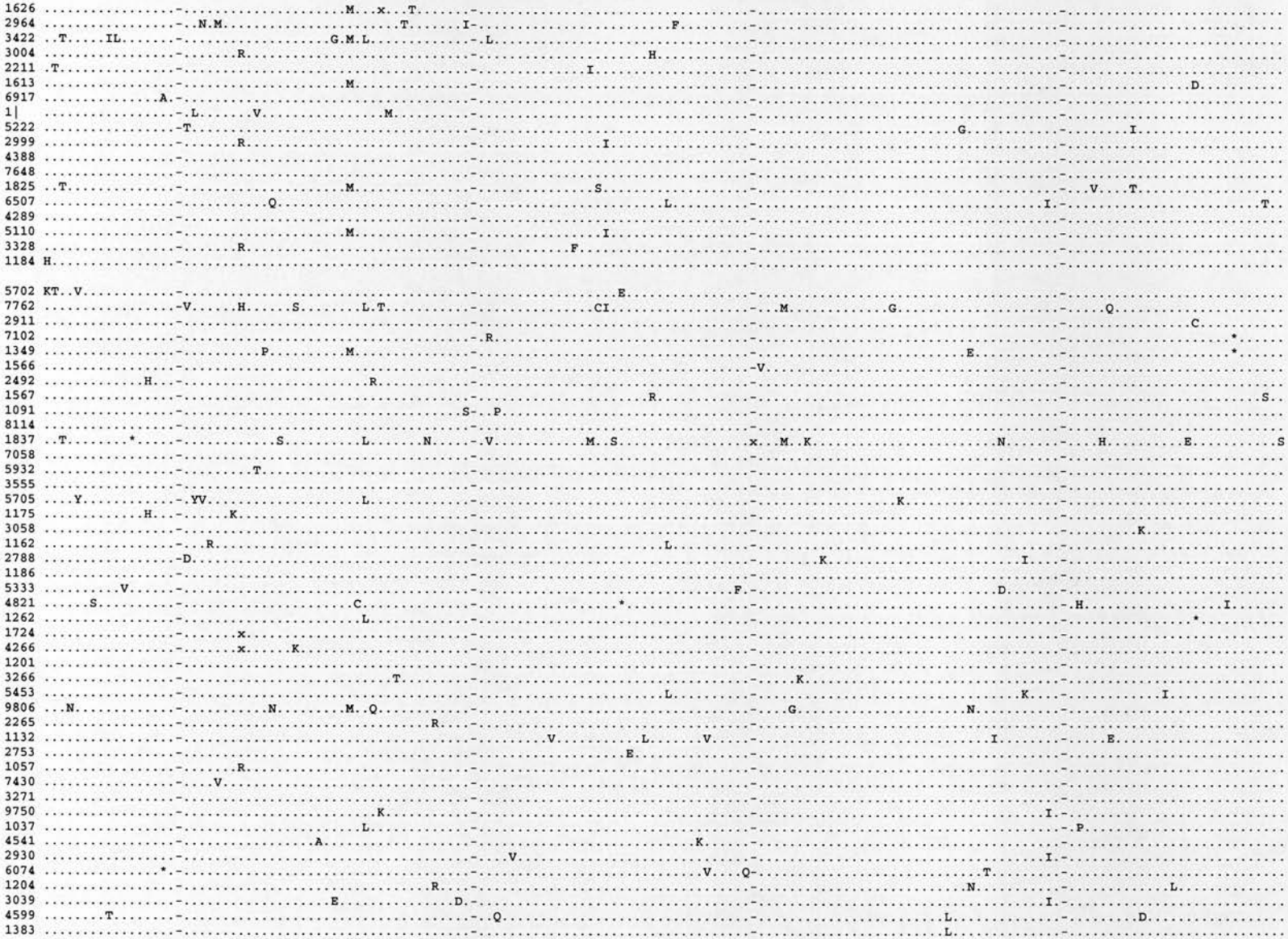


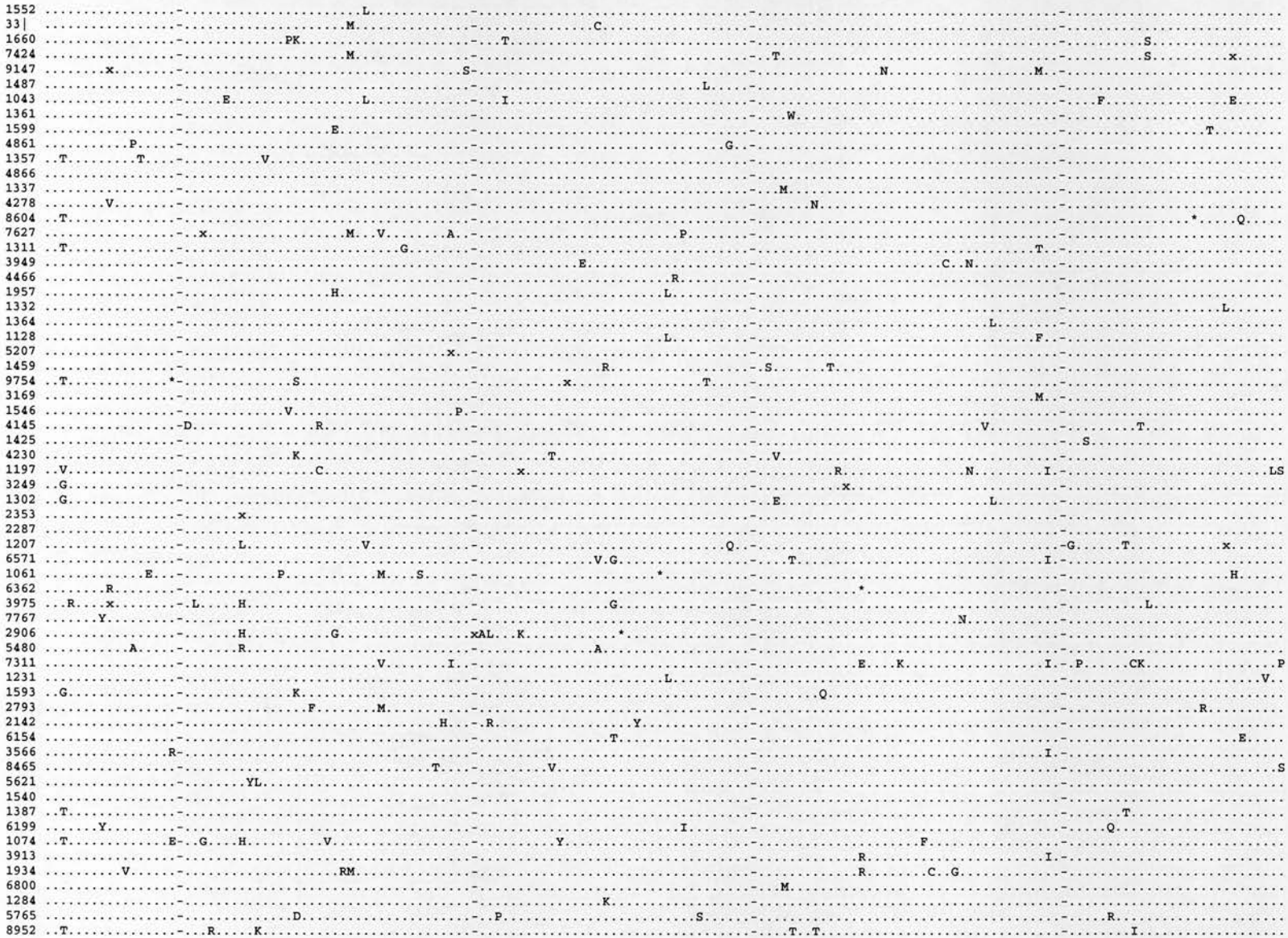

801
960
3111 FLDTYTLPRLNQEEVESLNRPITGSEIVAIINSLPTKKSPGPDGFTAEFYQ--RYKEELVPFLLKLFQSIEKEGILPNSFYEASIILIPKPGRDTTKK--ENFRPISLMNIDAKILNKILANRIQQHIKKLIHHDQVGFIPGMQGWFNIRKSIN-VIQHI

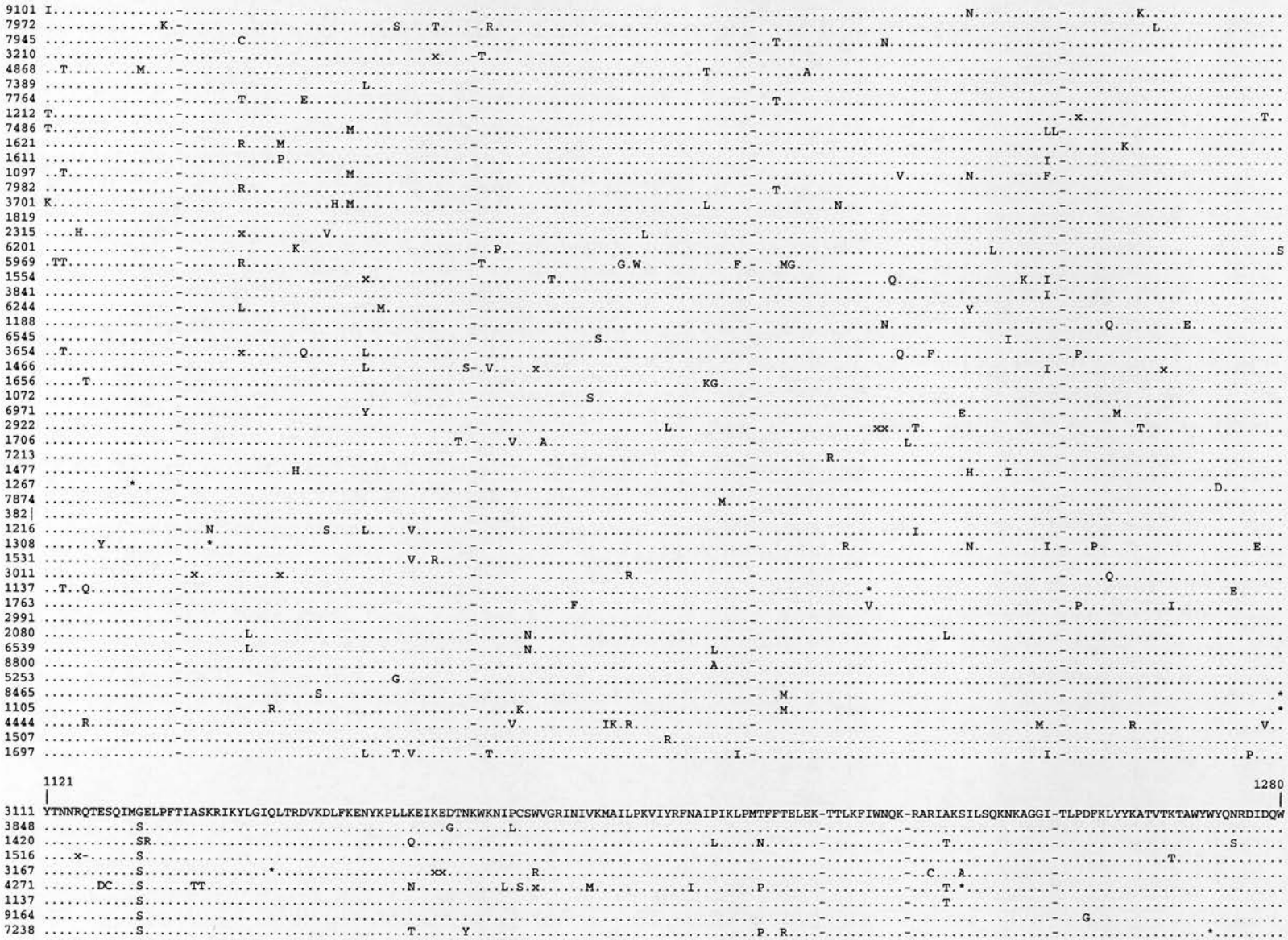


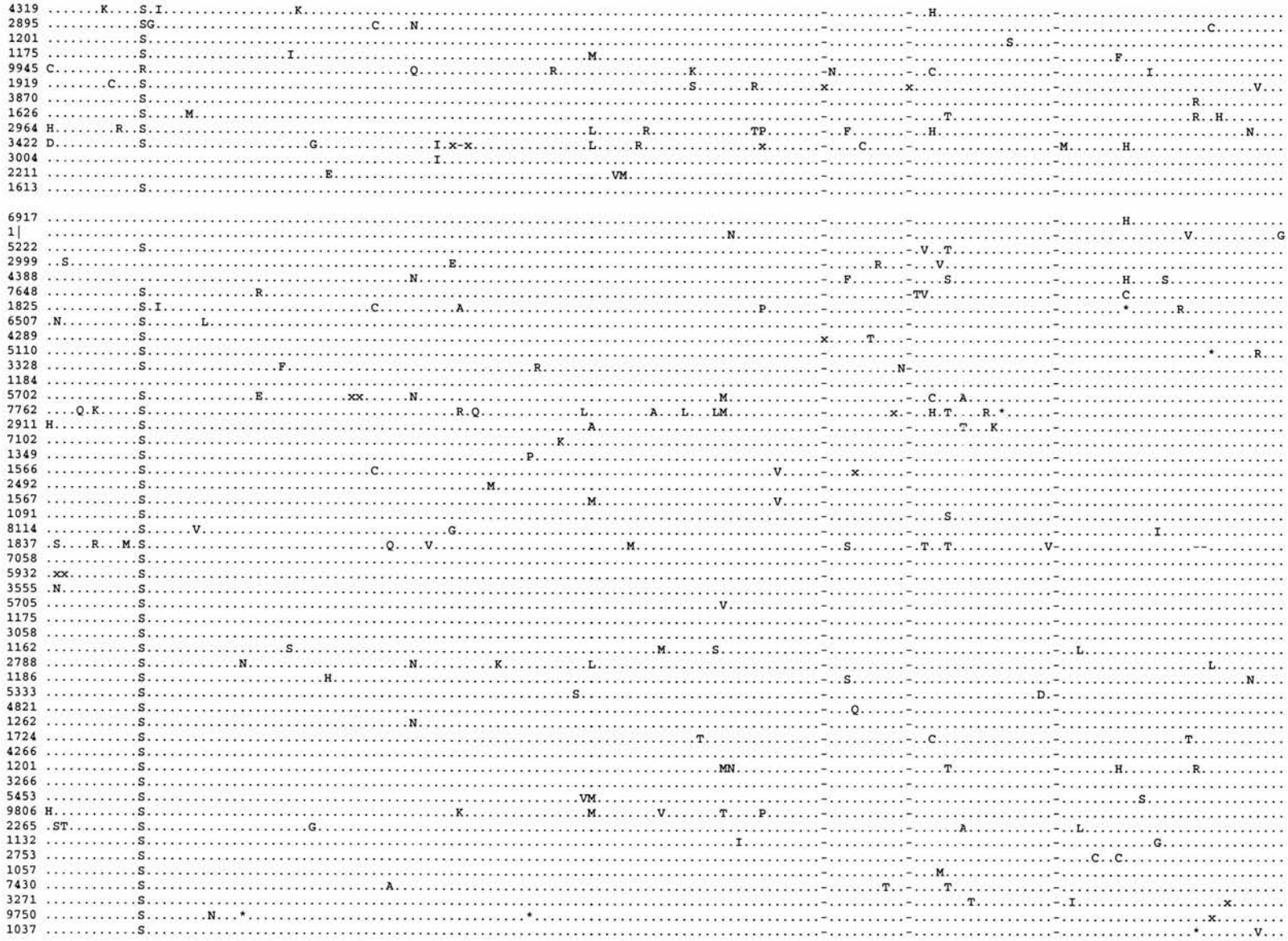


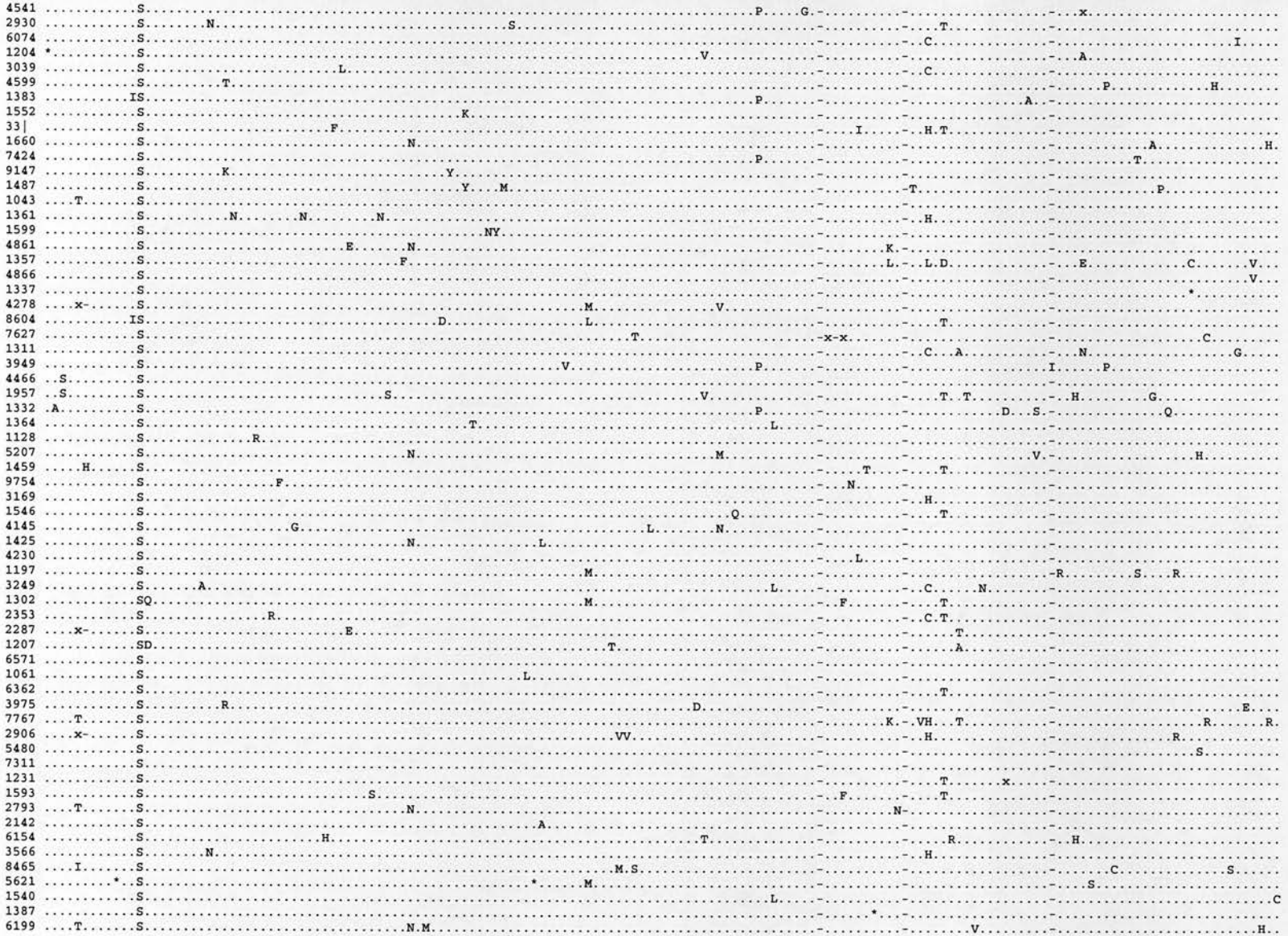

961
1120
NRAKDKNHMIISIDAEK-AFDKIQQPFMLKTLNKLGIDGTYFKIIRAIYDKPTAN-IILNGQKLEAFPLKTGTRQGCPLSPLLFNIVLEVL-ARAIRQEKEIKGIQLGKEEVKLSLFADDMIVYLENPIVS-AQNLLKLISNFSKVSGYKINVQKSQAFL

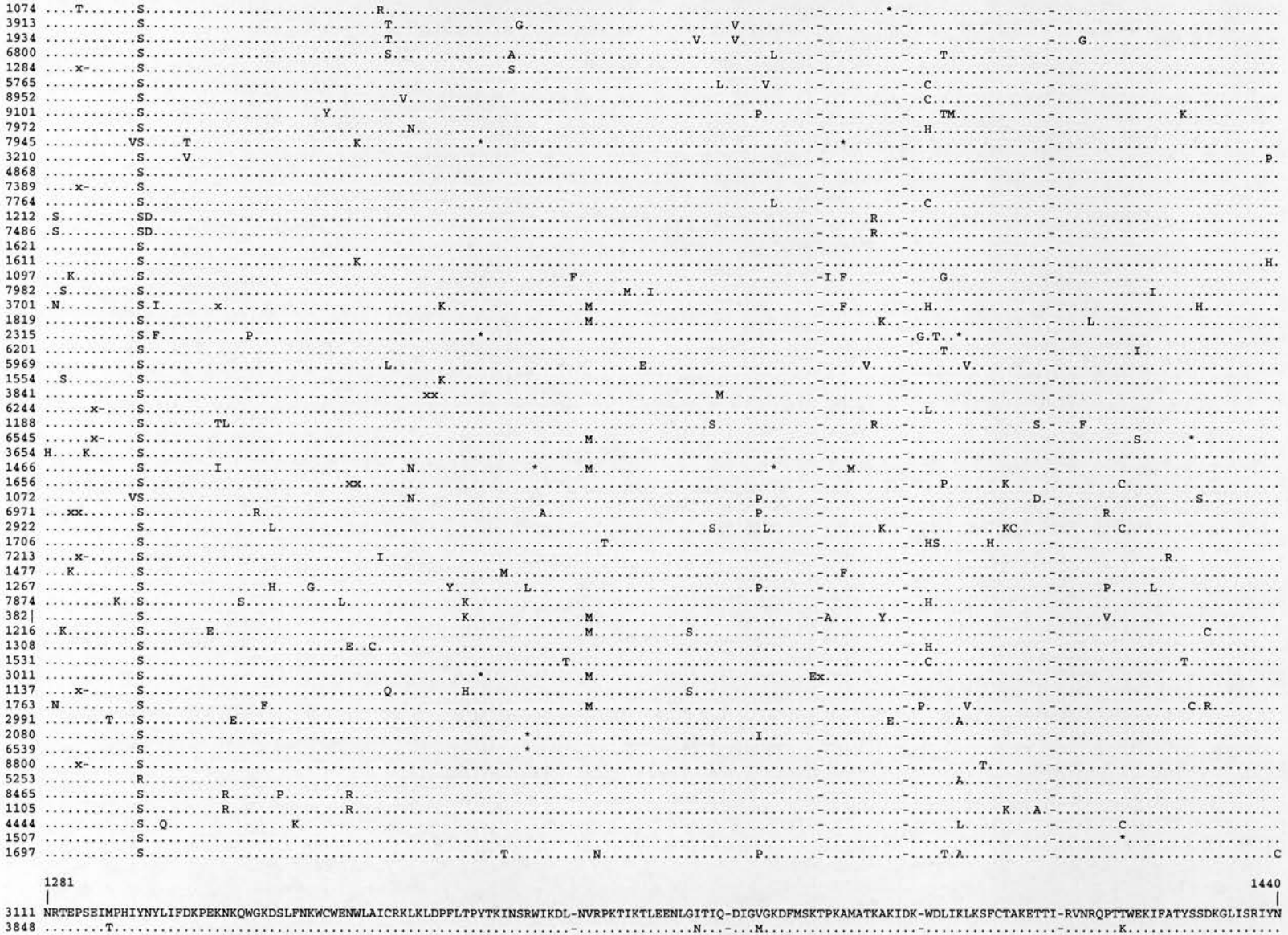


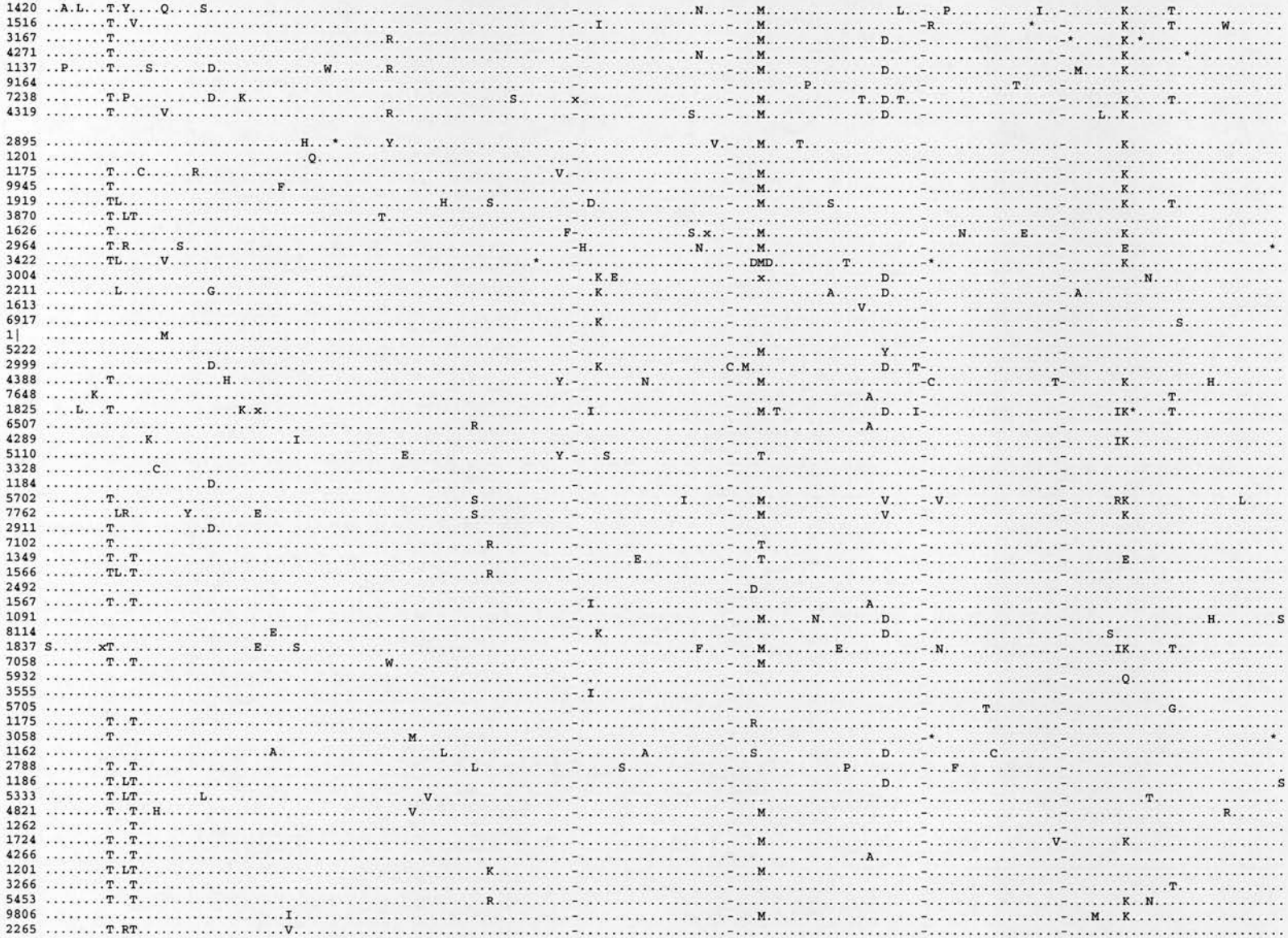


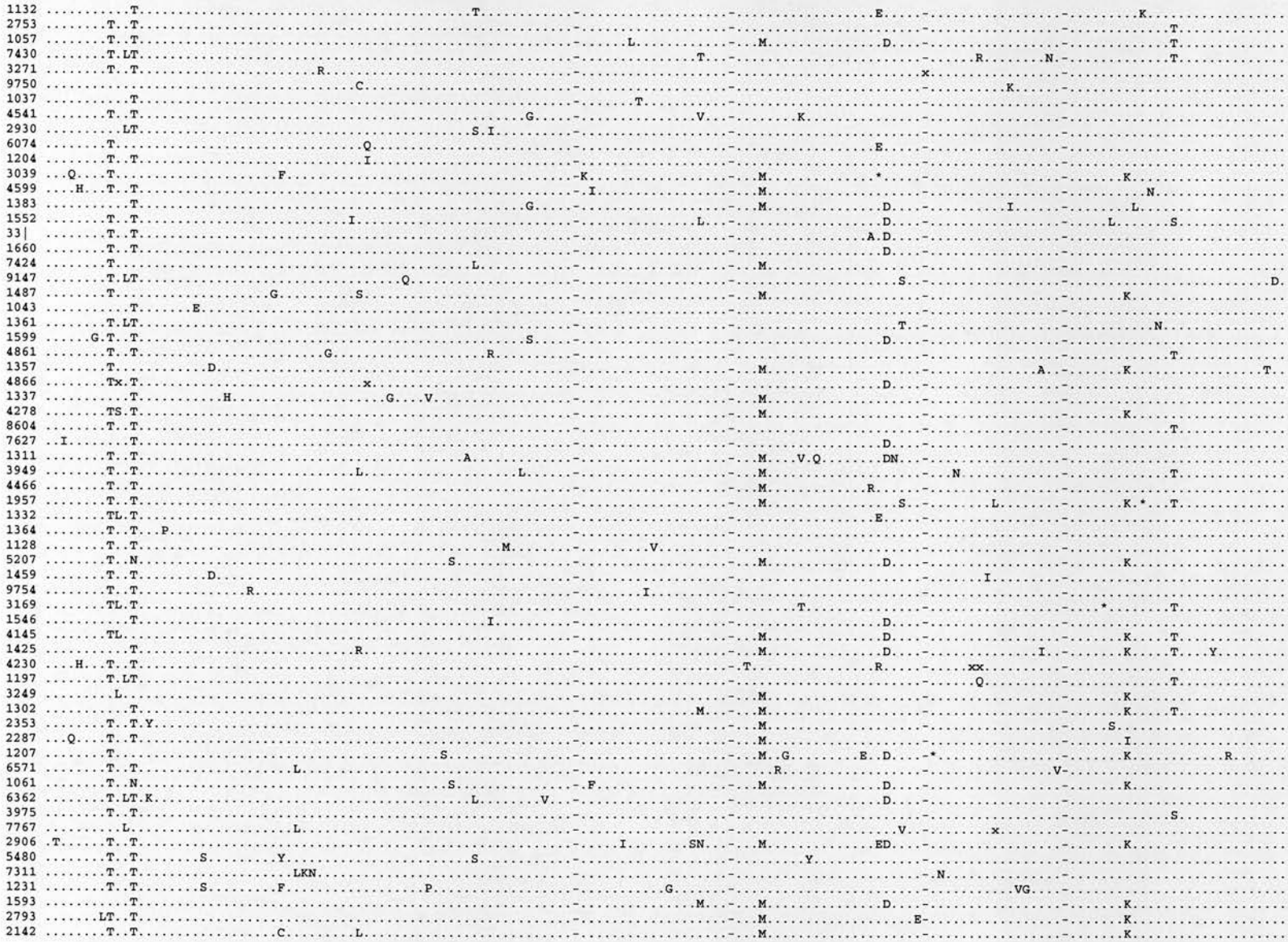

1121
1280
YTNNRQTESQIMGELPFTIASKRIKYLGIQLTRDVKDLFKENYKPLLKEIKEDTNKWKNIPCSWVGRINIVKMAILPKVIYRFNAIPIKLPMTFFTELEK-TTLKFIWNQK-RARIAKSILSQKNKAGGI-TLPDFKLYYKATVTKTAWYWYQNRDIDQW

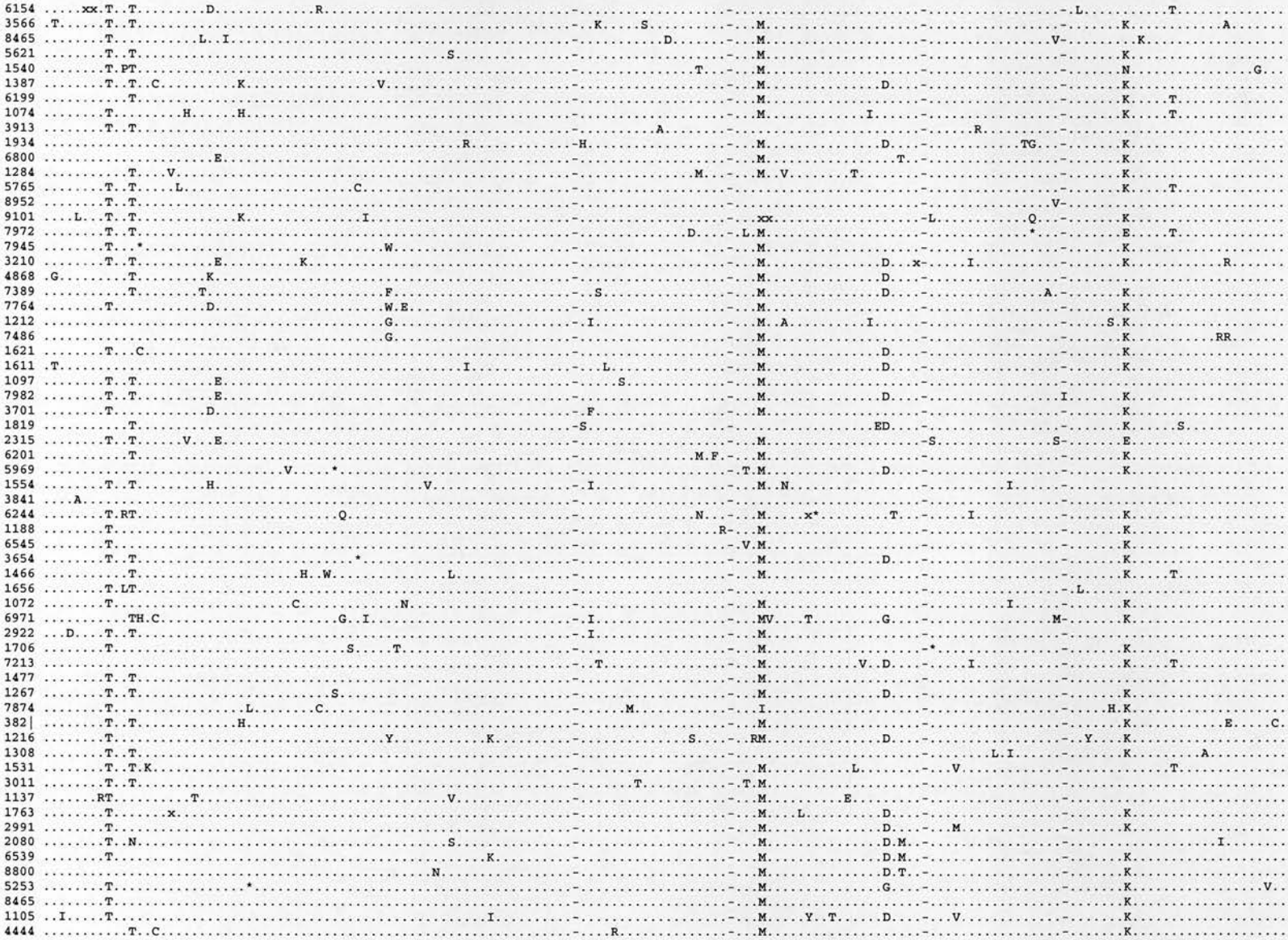


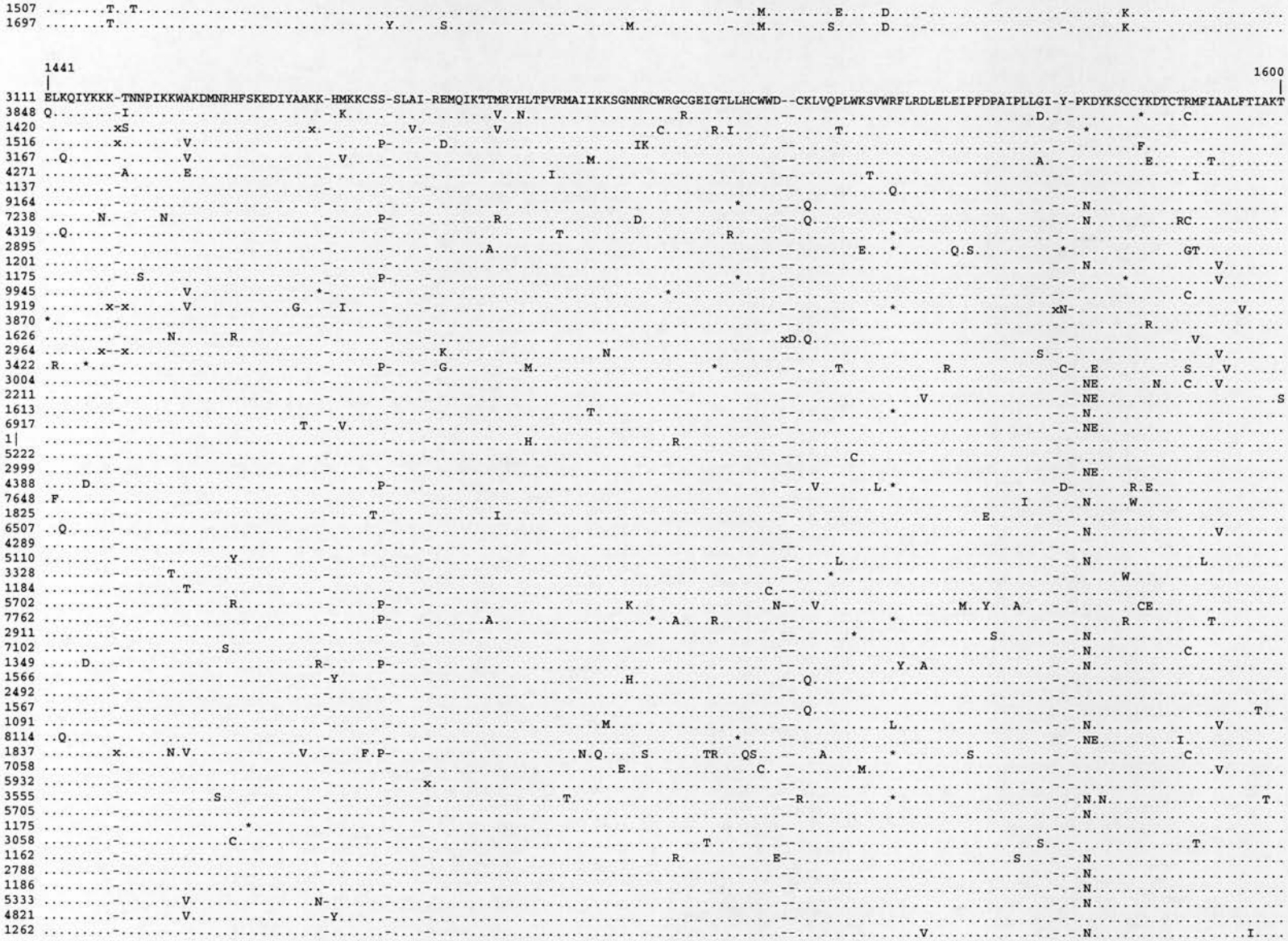


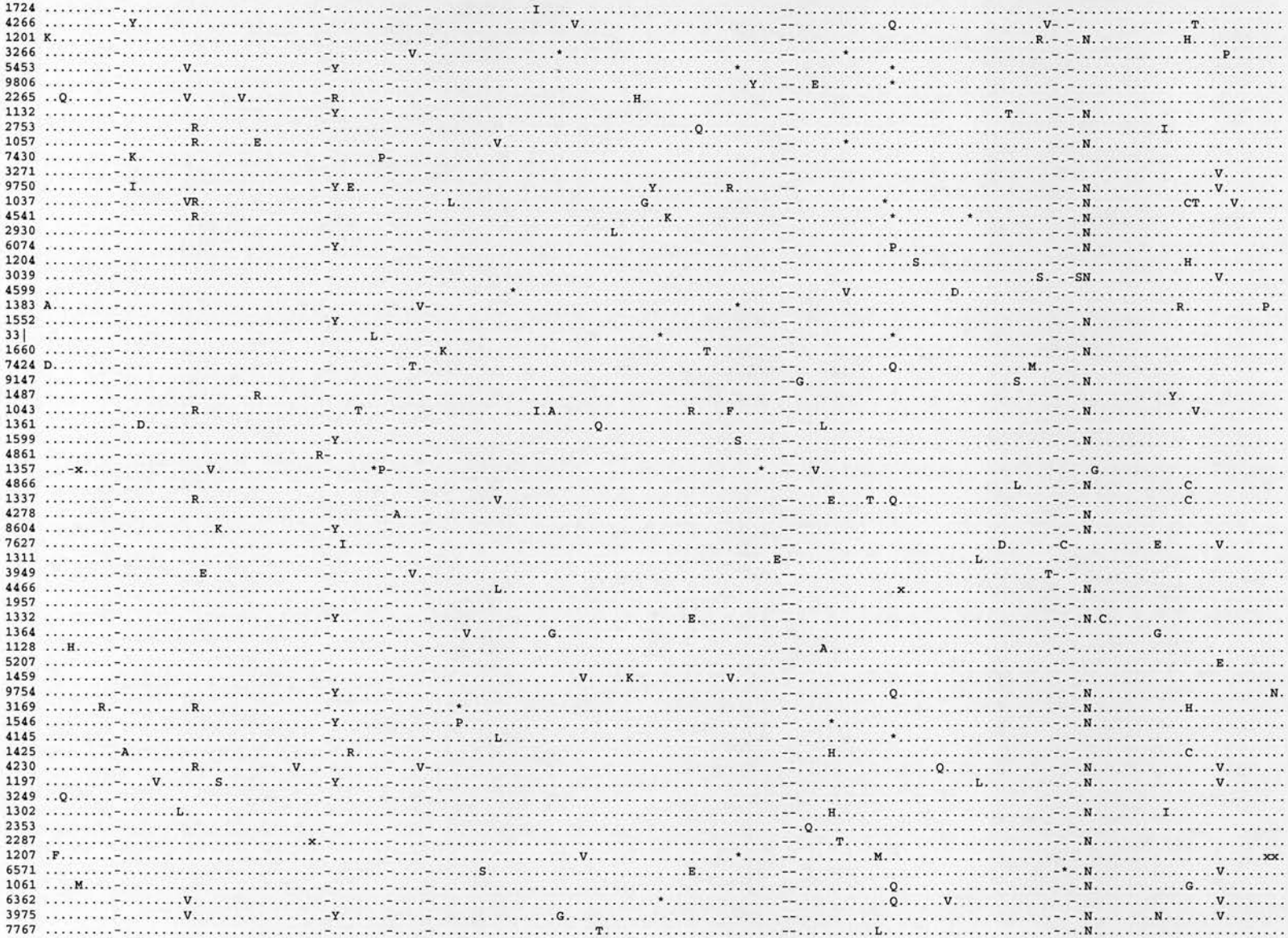


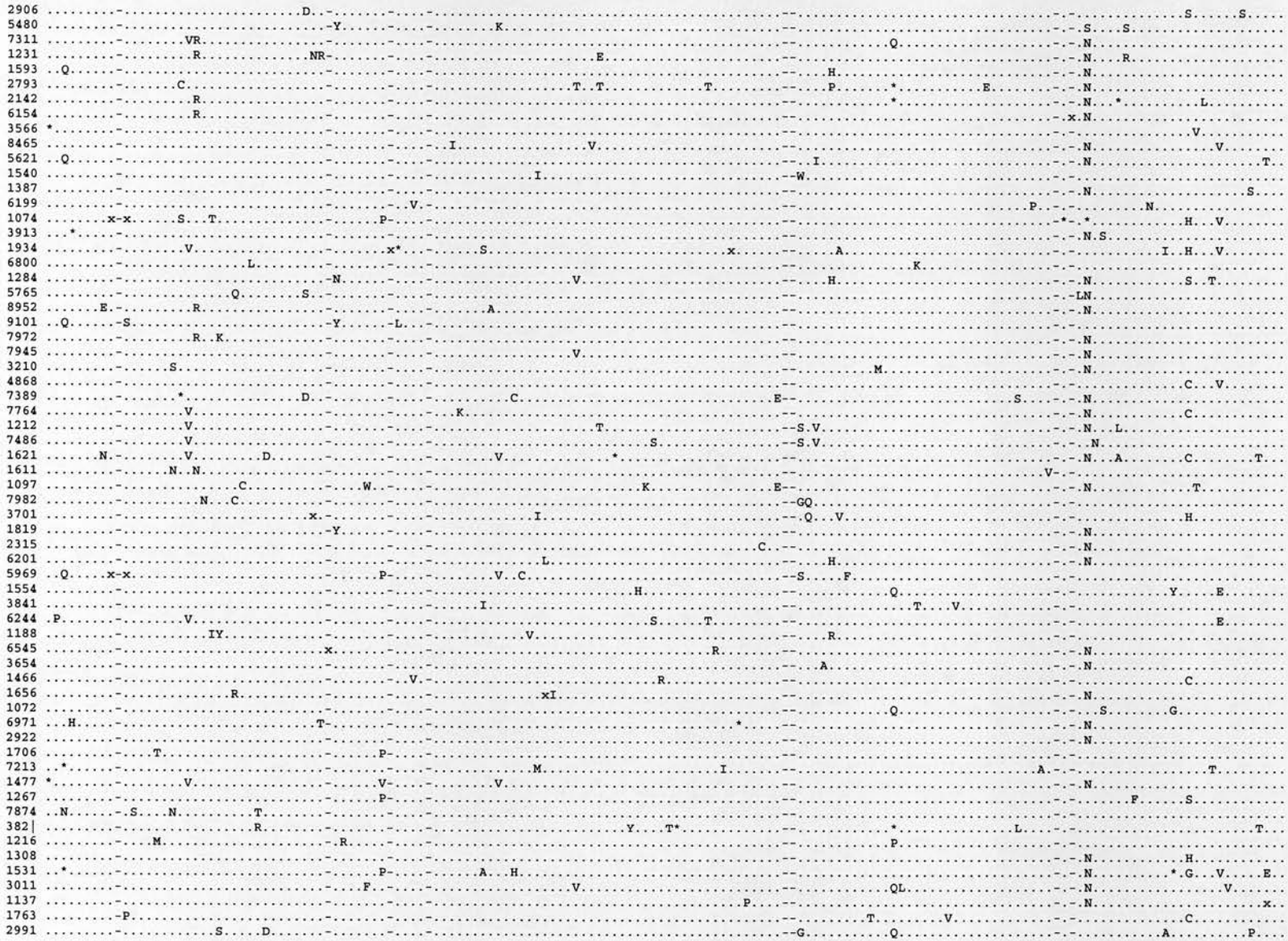


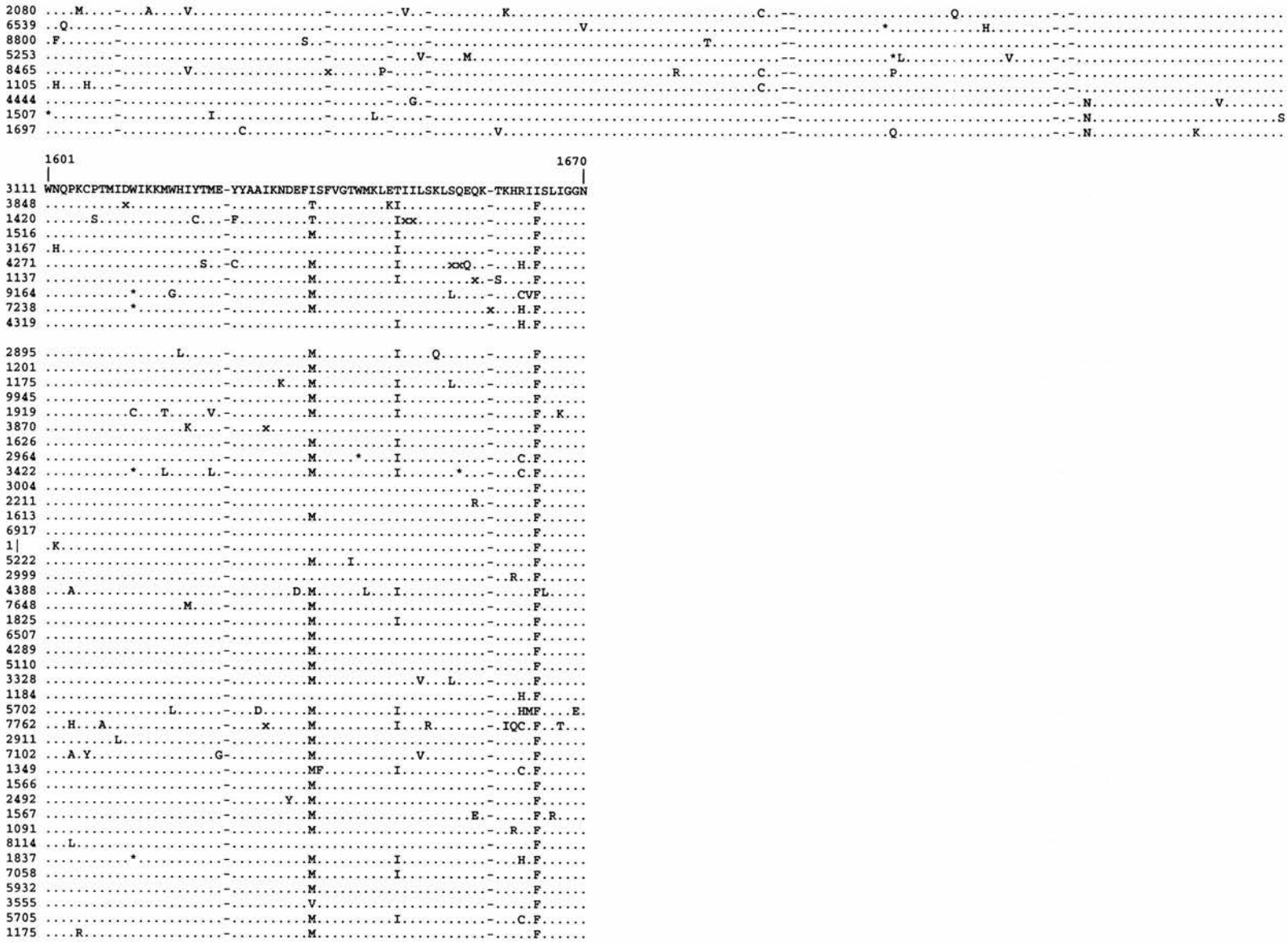


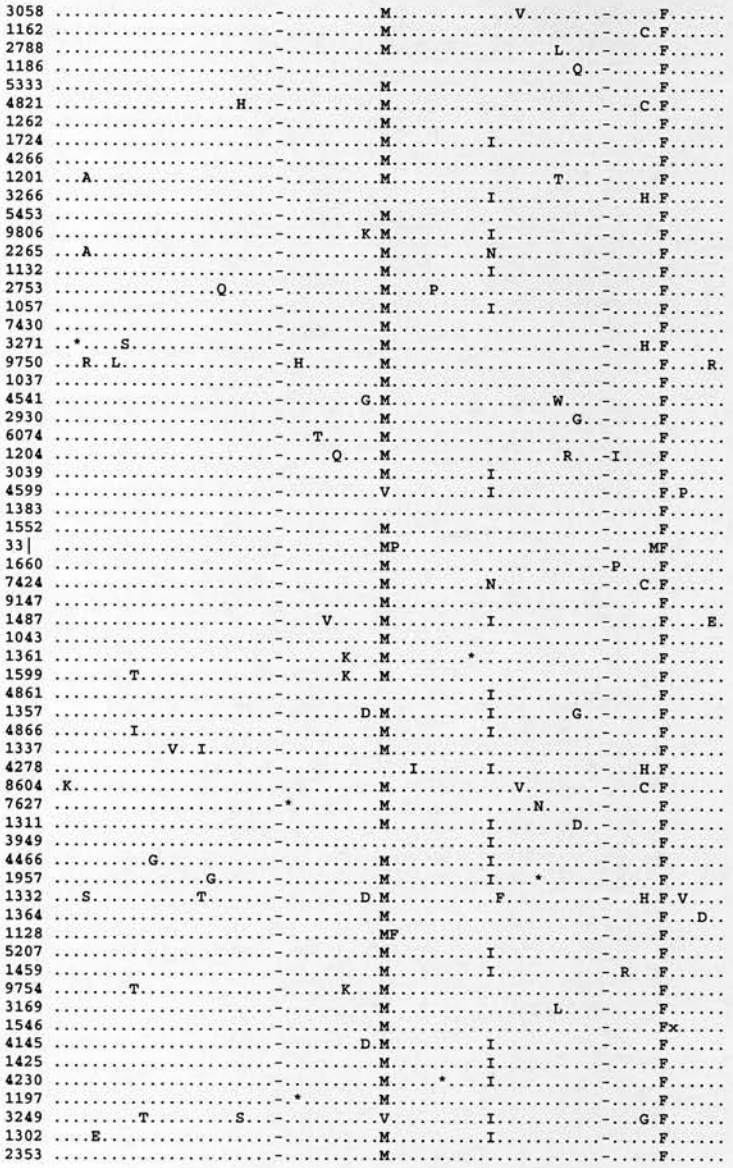


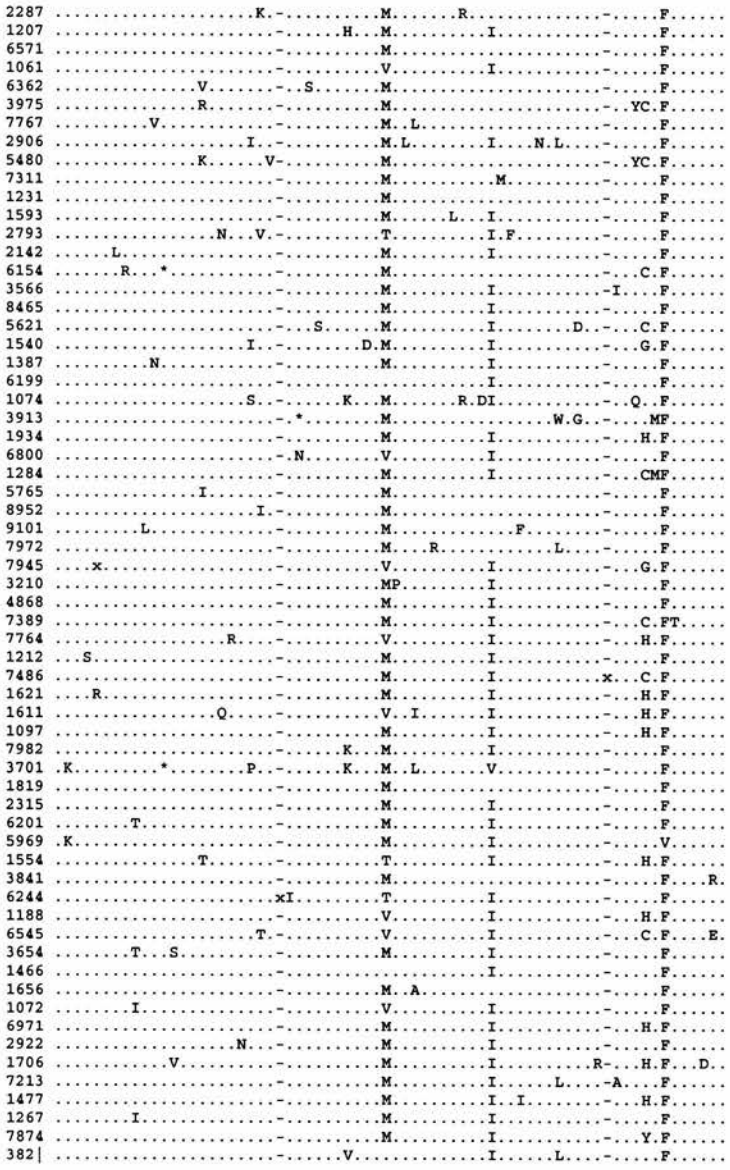

1441
1600
ELKQIYKKK-TNNPIKKWAKDMNRHFSKEDIYAAKK-HMKKCSS-SLAI-REMQIKTTMRYHLTPVRMAIIKKSGNNRCWRGCGEIGTLLHCWWD--CKLVQPLWKSVWRFLRDLELEIPFDPAIPLLGI-Y-PKDYKSCCYKDTCTRMFIAALFTIAKT











```
1216 ......................-..........M..........I..........-.....F......
1308 ................Y.....-..........M.......N.........L....-...H.F......
1531 ......................-..........M..........I..........-.....F......
3011 ...A.........Q.........-......K...M..........I..........-.....F......
1137 ....................I..-..........M..........I......L....-.....F......
1763 ......................-..........M..........I..........-...C.F......
2991 ...A..............N..-...G......M....A.....I...R.......-.....F......
2080 ....................N..-.........LM..........I..........-x.Y..F......
6539 ......................-....x.....M..........I..........-.....F......
8800 ........I.............-..........M..........I..........-...H.F......
5253 ......................-..........M..........I..........-...H.F......
8465 ......................-..........M..........I..........-.....F......
1105 ......................-.......E..M..........I..........-.....F......
4444 ......................-..........M..........I..........-....MF......
1507 ..............T.......-..........M..........I..........-.....F......
1697 ......................-..........M.....................-...C.F......
```

A.11 SVA Alignment of region which corresponds to *env* of HERV-K(HML-2)

```
           1                                 34
           |                                 |
envand3ltr HRERAMMTMAVLSKRKGGNVGKSKRDQIVTVSV*
 Homo sapi .WGW........WN..A.R...KL.N.M.AG..W
 Homo sapi .WGW........WN..A.R...KL.N.M.AG..W
 Homo sapi .WGW........WN..A.R...KL.NRM.AG..W
 Homo sapi .WGW.....V..WN..A.R...KL.NRM.AG..W
 Homo sapi .*...G......WN...EK...RL.NRM.A...*
 Homo sapi .WGW........WN..A.R...KL.NRM.AGF.W
 Homo sapi .WGW........WN..A.R...KL.N.M.AG..W
 Homo sapi .WGW........WN..A.R...KL.NRM.AG..W
 Homo sapi .WGW........WN..A.R...KL.NRM.AG..W
 Homo sapi .*...R....A.WN..A.K...RL.NRM.A..L*
 Homo sapi .WGW........WN..A.R...KL.NRM.A...W
 Homo sapi .WGW........WN..A.R...KL.NRM.AG..W
 Homo sapi .WGW........WN..A.RL..KL.NRM.AG..W
 Human chr .WGW........WN..A.R...KL.NRM.AG..W
 Human chr .*...........N..V.....R*.NR..A...*
NUMBER=1&s .WGW........WN..A.R...KL.NRM.AG..W
NUMBER=1&s .*...........NGx......R..SR......W
 Homo sapi .*.G.........N........R...R......L
 Homo sapi .W*W........WN..A.R...EL.NRMF.G..W
 Homo sapi .WGW........WN..A.R...KL.NRM.AG..W
 Homo sapi .WGW........WN..A.R...KL.NRM.AG..W
 Homo sapi .WGW........WN..A.R...KL.NRM.AG..W
 Homo sapi .*..........WN.......NR*.N...A...*
 Homo sapi .WGW........WN..A.R...KL.NLM.AG..W
 Homo sapi .*.W.........N........G*.N..LA...*
 Homo sapi .WGW........WN..A.R...KL.NRM.AG..W
 Homo sapi .WGW........WN..A.R...KL.NRM.AG..W
 Homo sapi .*...R..V.A.WNGEA.G...GL.NRM.A...*
 Homo sapi .WGW........WN..A.R...KL.NRM.AG..W
      886_ .WRW..V.....WN..A.R...KL.NRMGAG..W
 Homo sapi .WGW........WN..A.R...KL.NRM.AG..W
 Homo sapi .WGW........WN..A.R...KL.N.M.AG..W
```

```
Human chr .*..VR....A.WN..A.K...RL.NRM.A...*
Human DNA .WGW........WN..A.R...KL.N.M.AG..W
Homo sapi .WGW........WN..A.R...KL.NRM.AG.LW
Homo sapi .WGW........WN..A.R...KL.N.M.AGF.W
Human DNA .WGW........WN..A.R...KL.N.M.AGF.W
Homo sapi .WGW........WN..A.R...KL.N.M.AG..W
Homo sapi .WGW........WN..A.R...KL.NRM.AG..W
Human DNA .WGW........WN..A.R...KL.NRM.AG..W
Homo sapi .WGW........WN..A.R...KL.NRM.AG..W
Homo sapi .WGW........WN..A.R...KL.NRM.AG..W
Human DNA .WGW........WN..A.R...KL.NRM.AG..*
Homo sapi .*...........N........xx..R.....G*
Homo sapi .WGW........WN..A.R...KL.NRM.AG..W
Homo sapi .WGW........WN..A.R...KL.NRM.AG..W
Homo sapi .WG*........WN..A.R...KL.NRM.AG..W
Human DNA .WGW........WN..A.R...KL.NRM.AG..*
Human DNA .*..........WN....K...ILSN.M.A...W
Homo sapi .WG*........WN..A.R...KL.NRM.AG..W
Homo sapi .WGW........WN..A.R...KL.NRM.AG..W
Human DNA .WGW........WN..A.R...KL.NRM.AG..W
Homo sapi .WGW........WN..A.R...KL.NRM.AG..W
Homo sapi .WGW........WN..A.R...KL.N.M.AG..W
Homo sapi .*.G........WN....K...RL.NRM.AM..*
Homo sapi .WGW........WN..A.R...KL.NRM.AG..W
Homo sapi .WGW........WN..A.R...KL.NRM.AG..W
Homo sapi .WGW........WN..A.R...KL.N.M.AG..W
Human DNA .WGW........WN..A.R...RL.NRM.AG..W
Homo sapi .WGW........WN..A.R...KL.NRM.AG..W
Homo sapi .WGW........WN..A.R...KL.NRM.AG..W
Human DNA .WGW........WN..A.R...KL.NRM.AG..W
Homo sapi .WGW........WN..A.R...KL.NRM.AG..W
Homo sapi .*...........N........R*.N...A...*
Homo sapi .WGW........WN..A.R...KL.NRM.AG.LW
Homo sapi .*..........WN..R.K...RL.NRM.....*
Human DNA .WGW........WN..A.R...KL.NRM.AG..W
Homo sapi .WGW........WN..A.R...KL.NRM.AG..W
Homo sapi .*..........WN..A.R...KL.NRMIAG..W
```

```
 Homo sapi .WGW........WN..A.R...KL.NRM.AG..W
 Homo sapi .WGW........WN..A.R...KL.NRM.AG..W
 Homo sapi .WGW........WN.RA.R..RKL.NRM.AG..W
NUMBER=1&s .WGW........WN..A.R...KL.NRM.AG..W
      210_ .WGW........xNx.x.R...KL.N.M.AG..W
 Homo sapi .*D.........WN....K...RL.NRM.A...W
 Homo sapi .WGW........WN..A.R...KL.NRM.AG..W
 BAC seque .WGW........WN..A.R...KL.NRM.AG..W
NUMBER=1&s .*...........NGx......R..SR......W
 Homo sapi .*.G.........N........R...R......L
 Homo sapi .WGW........WN..A.R...KL.NRM.AG.LW
 Homo sapi .WGW........WN..A.R...KL.N.M.AGF.W
 Homo sapi .WGW........WN..A.R...KL.NRM.AG..W
 Human DNA .WGW........WN..A.R...KL.N.M.AG..W
 Homo sapi .WGW........WN..A.R...KL.NRM.AG..W
 Homo sapi .WGW........WN..A.R...KL.NRM.AG..W
 Human DNA .WGW........WN..A.R...KL.N.M.AG..W
 Human chr .WGW........WN..A.R...KL.NRM.AG..W
 Homo sapi .WGW........WN..A.R...KL.NRM.AG..W
 Homo sapi .WGW........WN..A.R...KL.NRM.AG..W
 Homo sapi .WGW........WN..A.R...KL.NRM.AG..W
 Homo sapi .WGW........WN..A.R...KL.N.M.AG..W
 Human chr .WGW........WN..A.R...KL.N.M.AG..W
 Homo sapi .WGW........WN..A.R...KL.NRM.AG..W
 Homo sapi .WGW........WN..A.R...KL.NRM.AG..W
 Homo sapi .WGW........WN..A.R...KL.NRM.AG..W
 Homo sapi .WGW........WN..A.R...KL.NRM.AG..W
 Homo sapi .WGW........WN..A.R...KL.NRM.AG..W
 Homo sapi .WGW........WN..A.R...KL.NRM.AG..W
 Homo sapi .WGW........WN..A.R...KL.NRM.AG..W
 Homo sapi .WGW........WN..A.R...KL.N.M.AG..W
```

APPENDIX B

Table of Contents

2.3 Polymerase Chain Reaction

Figure B.1 Chemicals Used

10 × TBE: 324g Tris base (Sigma)

85 g Boric acid (Sigma)

19 g EDTA (Sigma)

Make up to 2L with distilled water

1 x TBE used as electrophoresis buffer in gel tank

2.2.1 Collection of Genomic DNA

Table B.1 Anthropometric Data Accompanying the Buccal Swabs

samples									
Sample ID	Date of Birth	SEX	Individual Born	Mother Born	Father Born	Individual Ethnicity	Mother Ethnicity	Father Ethnicity	Further Comments
B01	31/01/1975	FEMALE	KOREA	KOREA	KOREA	KOREAN	KOREAN	KOREAN	
B02	17/07/1976	FEMALE	FRANCE	FRANCE	FRANCE	FRENCH	FRENCH	FRENCH	
B03	26/12/1981	FEMALE	NORWAY	NORWAY	NORWAY	NORWEGIAN	NORWEGIAN	NORWEGIAN	
B04	01/03/1973	MALE	PAKISTAN	PAKISTAN	PAKISTAN	ASIAN	ASIAN	ASIAN	
B05	01/01/2001	MALE	GERMANY	GERMANY	GERMANY	GERMAN	GERMAN	GERMAN	
C01	04/07/1965	FEMALE	CHINA (SHANGHAI)	CHINA (SHANGHAI)	CHINA (SHANGHAI)	CHINESE	CHINESE	CHINESE	
C02	07/05/1958	MALE	CHINA (BEJING)	CHINA (BEJING)	CHINA (BEJING)	CHINESE	CHINESE	CHINESE	
C03	13/12/1954	FEMALE	CHINA (HEBEI PROVINCE)	CHINA (HEBEI PROVINCE)	CHINA (HEBEI PROVINCE)	CHINESE	CHINESE	CHINESE	
C04	14/04/1958	MALE	CHINA (BEJING)	CHINA (HEBEI PROVINCE)	CHINA (LIAONING PROVINCE)	CHINESE	CHINESE	CHINESE	
C05	22/12/1952	MALE	CHINA (BEIJING)	CHINA (BEIJING)	CHINA (BEJING)	CHINESE	CHINESE	CHINESE	
C06	27/11/1956	MALE	CHINA (SHANGDUNG PROVINCE)	CHINA (SHANGDUNG PROVINCE)	CHINA (SHANGDUNG PROVINCE)	CHINESE	CHINESE	CHINESE	

samples									
Sample ID	Date of Birth	SEX	Individual Born	Mother Born	Father Born	Individual Ethnicity	Mother Ethnicity	Father Ethnicity	Further Comments
C07	15/07/1980	FEMALE	CHINA (SICHUAN PROVINCE)	CHINA (SICHUAN PROVINCE)	CHINA (SICHUAN PROVINCE)	CHINESE	CHINESE	CHINESE	
C08	29/03/1959	MALE	CHINA (BEJING)	CHINA (ZHE JIANG PROVICE)	CHINA (SHANDUNG PROVINCE)	CHINESE	CHINSE	CHINESE	
C09	18/12/1953	MALE	CHINA (BEIJING)	CHINA (SHANDUNG PROVINCE)	CHINA (SHANGDUNG PROVINCE)	CHINESE	CHINESE	CHINESE	
C10	09/10/1955	FEMALE	CHINA (BEIJING)	CHINA (HEIBEI PROVINCE)	CHINA (HEBEI PROVINCE)	CHINESE	CHINESE	CHINESE	
D01	11/09/1989	MALE	JARKARTA	WEST JAVA	KALIMANTAN TIMUR	JAVANESE	JAVANESE	BORNEO / DAYAKNESE	
D02	13/01/2002	FEMALE	JAKARTA	YOGYAKARTKA	YOGYAKARTKA	JAVANESE	JAVANESE	JAVANESE	
D03	01/01/2001	FEMALE	ACEH	ACEH	ACEH	ACEHNESE	ACEHNESE	ACEHNESE	
D04	12/08/1972	FEMALE	JAKARTA	WEST JAVA	EAST JAVA	JAVANESE	JAVANESE	JAVANESE	
D05	05/05/1986	FEMALE	MADURA	MADURA	MADURA	MADURANESE	MADURANESE	MADURANESE	
D06	26/03/1969	FEMALE	BANDUNG	N.SUMATRA	N.SUMATRA	BATAK	BATAK	BATAK	
D07	01/02/1953	FEMALE	JAKARTA	YOGYAKARTA	MAGELANG	JAVANESE	JAVANESE	JAVANESE	
K01	01/01/2001	FEMALE	HUNGRY	HUNGRY	HUNGRY	HUNGARIAN	HUNGARIAN	HUNGARIAN	
K02	01/01/2001	MALE	PAKISTAN	PAKISTAN	PAKISTAN	PAKISTANI	PAKISTANI	PAKISTANI	
K03	01/01/2001	FEMALE	CHINA	CHINA	CHINA	CHINESE	CHINESE	CHINESE	
K04	01/01/2001	MALE	GREECE	GREECE	GREECE	GREEK	GREEK	GREEK	
K05	01/01/2001	MALE	CHINA (KALIMANTAN)	CHINA	CHINA	CHINESE	CHINESE	CHINESE	
K06	01/01/2001	MALE	W.CHINA	W.CHINA	W.CHINA	CHINESE	CHINESE	CHINESE	

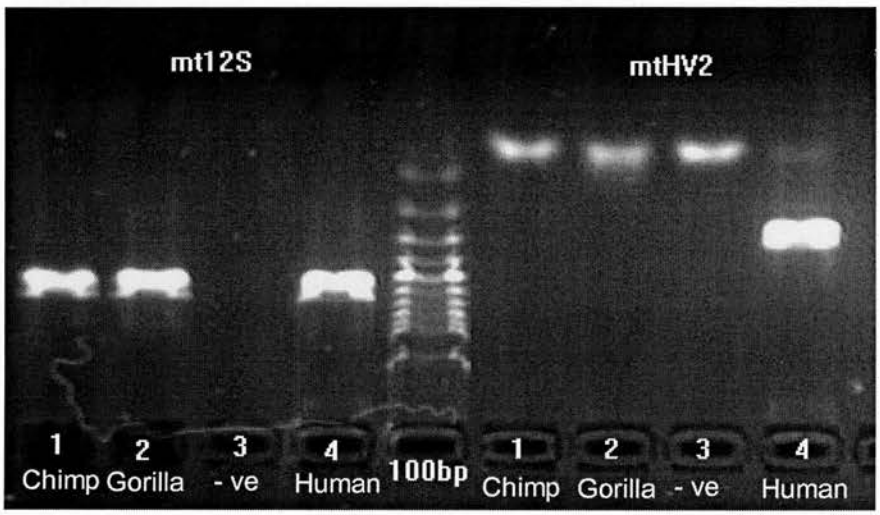
samples									
Sample ID	Date of Birth	SEX	Individual Born	Mother Born	Father Born	Individual Ethnicity	Mother Ethnicity	Father Ethnicity	Further Comments
K07	01/01/2001	MALE	W.JAPAN	W.JAPAN	W.JAPAN	JAPANESE	JAPANESE	JAPANESE	
K08	01/01/2001	MALE	GERMANY	GERMANY	GERMANY	GERMAN	GERMAN	GERMAN	
K09	01/01/2001	FEMALE	CHINA	CHINA	CHINA	CHINESE	CHINESE	CHINESE	
K10	01/01/2001	MALE	NEW YORK	NEW YORK	NEW YORK	EASTERN EURO JEW	EASTERN EURO JEW	EASTERN EURO JEW	
K11	01/01/2001	FEMALE	ARGENTINA	ARGENTIAN	ARGENTINA	ARGINTINIAN	ARGINTINIAN	ARGINTINIAN	
K12	01/01/2001	FEMALE	GREECE	GREECE	GREECE	GREEK	GREEK	GREEK	
K13	01/01/2001	FEMALE	SPAIN	SPAIN	SPAIN	CATALAN SPANISH	CATALAN SPANISH	CATALAN SPANISH	
K14	01/01/2001	MALE	SPAIN (MADRID)	SPAIN	SPAIN	SPANIHS	SPANISH	SPANISH	
K15	01/01/2001	FEMALE	CHINA	CHINA	CHINA	CHINESE	CHINESE	CHINESE	
K16	01/01/2001	MALE	GERMANY	GERMANY	GERMANY	GERMAN	GERMAN	GERMAN	
K17	01/01/2001	FEMALE	CHINA	CHINA	CHINA	CHINESE	CHINESE	CHINESE	
K18	01/01/2001	FEMALE	SWITZERLAND	SWITZERLAND	SWITZERLAND	SWISS	SWISS	SWISS	
NO06	05/11/1965	MALE	PHOENIX ARIZONA			AMERICAN	QUECHUA INDIAN	NORTHERN EUROPEAN	INDIVIDUAL 50% INDIAN Father: Dutch and German
NO11		MALE	TAHTEQUAH OKLAHOMA	HANNA OKLAHOMA	HANNA OKLAHOMA	CREEK INDIAN	CREEK INDIAN	CREEK INDIAN	
NO12	29/03/1948	FEMALE	YUBA CITY CALIFORNIA	BONHAM TEXAS	CLOUD CHIEF OKLAHOMA	CHOCKTAW / GAELIC	GAELIC	CHOCKTAW	INDIVIDUAL 50% INDIAN
P1	17/01/1979	FEMALE	N.ENGLAND	N.ENGLAND	N.ENGLAND	BRITISH	BRITISH	BRITISH	
P2	21/10/1976	FEMALE	SWEDEN	SWEDEN	SWEDEN	SWEDISH	SWEDISH	SWEDISH	

samples									
Sample ID	Date of Birth	SEX	Individual Born	Mother Born	Father Born	Individual Ethnicity	Mother Ethnicity	Father Ethnicity	Further Comments
P3	09/06/1969	MALE	JAPAN	JAPAN	JAPAN	JAPANESE	JAPANESE	JAPENESE	
P4	30/04/1979	FEMALE	N.ENGLAND	N.ENGLAND	N.ENGLAND	ENGLISH	ENGLISH	ENGLISH	
P5	21/01/1977	FEMALE	N.ENGLAND	N.ENGLAND	N.ENGLAND	ENGLISH	ENGLISH	ENGLISH	
P6	10/01/1978	FEMALE	N.ENGLAND	N.ENGLAND	N.ENGLAND	ENGLISH	ENGLISH	ENGLISH	
P7	22/05/1976	FEMALE	N.ENGLAND	N.ENGLAND	N.ENGLAND	ENGLISH	ENGLISH	ENGLISH	
P8	10/09/1976	FEMALE	MALAYSIA	MALAYSIA	MALASIA	CHINESE	CHINESE	CHINESE	
SA01	25/06/1950	MALE	NATAL	NATAL	NATAL	ZULU	ZULU	ZULU	
SA02	24/09/1944	MALE	NATAL	NATAL	NATAL	ZULU	ZULU	ZULU	
SA03	20/04/1971	MALE	TRANSKES	TRANSKES	TRANSKES	XHOSA	XHOSA	XHOSA	
SA04	10/06/1972	MALE	NATAL	NATAL	NATAL	INDIAN	INDIAN	INDIAN	
SA05	04/09/1973	MALE	PRETORIA	PRETORIA	PRETORIA	TSWANA	TSWANA	TSWANA	
SA06	13/12/1977		NATAL	NATAL	NATAL	ZULU	ZULU	ZULU	
SA07	11/07/1971	MALE	NATAL (DURBAN)	NATAL (LADYSMITH)	NATAL)LADYSMITH)	ZULU	ZULU	ZULU	
SA08	11/11/1980		LUSIKISIICI (EASTERN CAPE)	NDESI	NDESI	XHOSA	XHOSA	XHOSA	
SA09	26/11/1976	Female	NATAL	NATAL	NATAL	ZULU	ZULU	ZULU	
SA10	20/07/1972	FEMALE	NATAL	NATAL	NATAL	INDIAN	INDIAN	INDIAN	
SA11	11/12/1970	MALE	NATAL (DURBAN)	NATAL (DURBAN)	NATAL (DURBAN)	INDIAN	INDIAN	INDIAN	
SA12	16/12/1971	MALE	NATAL (DURBAN)	NATAL (DURBAN)	NATAL (DURBAN)	INDIAN	INDIAN	INDIAN	
SA13	13/06/1963	MALE	NATAL	NATAL	NATAL	INDIAN	INDIAN	INDIAN	
SA14	24/06/1958	MALE	DURBAN	TRANSKEI	SWAZILAND	ZULU	XHOSA	SWAZI	

samples									
Sample ID	Date of Birth	SEX	Individual Born	Mother Born	Father Born	Individual Ethnicity	Mother Ethnicity	Father Ethnicity	Further Comments
				(EASTERN CAPE)					
SA15	12/12/1972	MALE	NATAL	NATAL	NATAL	ZULU	ZULU	ZULU	
SA16	17/10/1955	MALE	NATAL	NATAL	NATAL	ZULU	ZULU	ZULU	
SA17	30/04/1978	MALE	NATAL	NATAL	NATAL	ZULU	ZULU	ZULU	
SA18	28/03/1978	MALE	NATAL	NATAL	NATAL	ZULU	ZULU	ZULU	
SA19	22/09/1970	MALE	NATAL	NATAL	NATAL	ZULU	ZULU	ZULU	
SA20	24/02/1974	MALE	NATAL	NATAL	NATAL	ZULU	ZULU	ZULU	
SA21	28/08/1959	MALE	NATAL	NATAL	NATAL	ZULU	ZULU	ZULU	
SA22	05/11/1950		NATAL	NATAL	NATAL	ZULU	ZULU	ZULU	
SA23	29/01/1959	MALE	NATAL	NATAL	NATAL	ZULU	ZULU	ZULU	
SA24	12/09/1962	FEMALE	NATAL	NATAL	NATAL	ZULU	ZULU	ZULU	
V01	10/08/1978	FEMALE	HAWAII	HAWAII	HAWAII	HAWAIIAN	JAPANESE	JAPANESE, CHINESE	MOSTLY JAPANESE
V02	02/05/1980	MALE	LONDON	IRAQ	IRAQ	IRAQI JEW	IRAQI JEW	IRAQI JEW	
V03	19/09/1973	FEMALE	RUSSIA (MOSCOW)	RUSSIA (MOSCOW)	GERMANY (BERLIN)	RUSSIAN	RUSSIAN	RUSSIAN	
V04	24/02/1965	MALE	RUSSIA (KURSK)	RUSSIA (KURSK)	RUSSIA (MOSCOW)	RUSSIAN	RUSSIAN	RUSSIAN	
V06	08/04/1967	FEMALE	UKRAINE (CHARKOV)	UKRAINE (CHARKOV)	UKRAINE (CHARKOV)	RUSSIAN	RUSSIAN	UKRAINIAN	
V07	28/08/1951	FEMALE	SIERRA LEONE	SIERRA LEONE	SIERRA LEONE	AFRICAN	AFRICAN	AFRICAN	

3.2.3 Relative Age of HERV-K Proviruses

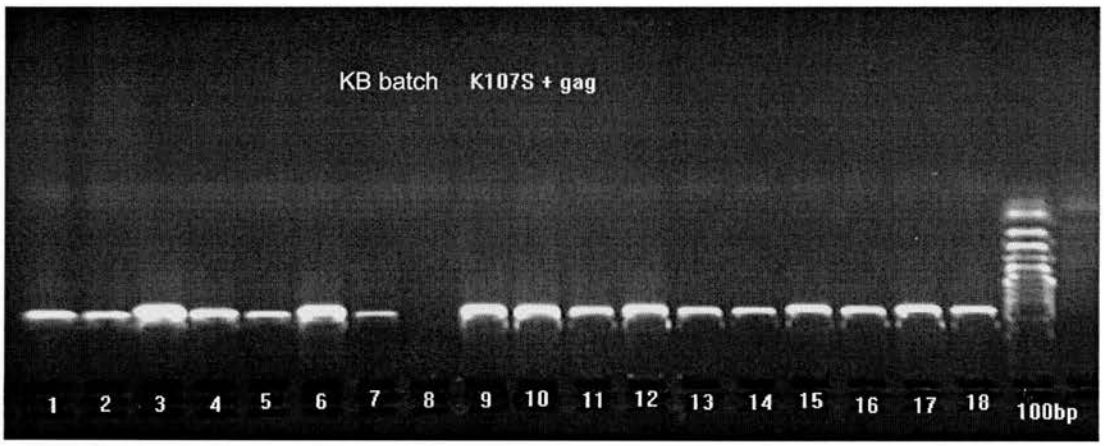
Figure B.2 Representative results for the PCR amplification of mitochondrial DNA in chimpanzee, gorilla and human samples.. Lanes 3 contain a negative control. The first set of 4 lanes show the amplification of the 12S rRNA region of the higher primate mitochondrial genome. The second set of lanes show the amplification for a 299 bp region of the second hypervariable segment (HV2) of the human mitochondrial genome. The lack of product in Lanes 1 and 2 indicates that the primate DNA samples are not contaminated with human DNA.



3.2.3 Relative Age of HERV-K Proviruses

Figure B.3 Sequence alignment of the PCR products generated using primers designed to amplify the 12S rRNA region in higher primate mitochondria.

Figure B.4 Representative results for the PCR determination of allelic variants of the HERV-K107 locus in European DNA samples. Lane 8 contains a negative control. These results indicate that all individuals possess at least one complete copy of the HERV-K107 provirus.



5.2.1 Comparison of Direct Repeats

Table B.2 Consistent Direct Repeat sequences of HERV-K(HML-2) Proviruses. [a] Relative age determined by the presence or absence of the provirus in primate species (Section 3.2.3). Branching dates from humans; 55 Mya for Prosimians, 45 Mya for New World Monkeys, 28 for Old World Monkeys, 20 Mya for Gibbons, 14 Mya for Orang-utans, 8.5 Mya for Gorillas and 6.3 Mya for Chimpanzee (Figure 3.8). U – Relative age undetermined.

(HML-2) Provirus	Age [a] (Mya)	Accession	5'Direct Repeat	3'Direct Repeat
K113	< 2	AY037928	CTCTAT	CTCTAT
K115	< 2	AC130464	CCTTT	CCTTT
K101	< 6.3	AC007326	ACCCAG	ACCCAG
K102	< 6.3	AL353807	GGGATG	GGGATG
K103	< 6.3	AL591164	ATGGGG	ATGGGG
K104	< 6.3	AC025757	CAGAAC	CAGAAC
K106	< 6.3	AC078785	GGCTGG	GGCTGG
K107	< 6.3	AC016577	ACTGC	ACTGC
K108	< 6.3	AC0104060	GGTTTC	GGTTTC
K109	< 6.3	AL590785	ATATGC	ATATGC
12q14.1	< 6.3	AC0025420	TGGTA	TGGTA
11q22.1	< 6.3	AP007776	TTGTG	TTGTG
HERV-K(II)	< 6.3	AC092902	GGCCC	GGCCC
3q27.2	< 6.3	AC069420	GGTACA	GGTACA
1p31.1	< 6.3	AC093156	ATGGA	ATGGA
K105	< 14	AF16419	GAATTC	GAATTC
K110	< 14	AC068728	TGAGAC	TGAGAC
11q23.2	< 14	AP000831	AGCCT	AGCCT
10p14	< 14	AC015686	CATTC	CATTC
HERV-K(I)	< 14	AC084198	GAGGT	GAGGT
4p16	U	AC105916	ATTTG	ATTTG

5.2.1 Comparison of Direct Repeats

Table B.3 Consistent Direct Repeat sequences of HERV-K(HML-2) Solitary LTRs. [a] Relative age determined by the presence or absence of the provirus in primate species (Section 3.2.2). Branching dates from humans; 55 Mya for Prosimians, 45 Mya for New World Monkeys, 28 for Old World Monkeys, 20 Mya for Gibbons, 14 Mya for Orang-utans, 8.5 Mya for Gorillas and 6.3 Mya for Chimpanzee (Figure 3.8).

HML-2 Solitary LTR	Age [a] (Mya)	Accession	5'Direct Repeat	3'Direct Repeat
1p22.1	< 6.3	AF370125	CAATTA	CAATTA
1p31.2	< 6.3	AL356736	GAAGG	GAAGG
1q22	< 6.3	AL135927	GGCAC	GGCAC
2p22.2	< 6.3	AC007390	TCACAG	TCACAG
2p23.14	< 6.3	AC021294	CTTCCA	CTTCCA
2p23.3	< 6.3	AC074117	AGAGG	AGAGG
2q33.2	< 6.3	AC074019	ATAGTC	ATAGTC
3p12.3	< 6.3	AF042089	ATCAG	ATCAG
3p21.31a	< 6.3	Z84493	TCCTG	TCCTG
3p21.31b	< 6.3	AC025548	GGTAAG	GGTAAG
3q26.31	< 6.3	AC068566	ATTAAT	ATTAAT
3q28	< 6.3	AC062008	AAGAG	AAGAG
4q13.3	< 6.3	AC055844	CATAAT	CATAAT
5p15.31	< 6.3	AC091985	CTAAAT	CTAAAT
5q23.1	< 6.3	AC010267	TTTTC	TTTTC
5q35.1	< 6.3	AC008648	GCAAG	GCAAC
6q15	< 6.3	AL021774	GAAGA	GAAGA
6q23.2	< 6.3	AL596188	CTGCTT	CTGCTT
6p21.32a	< 6.3	Z80898	ACTTC	ACTTC
6p21.32b	< 6.3	AC022567	ACCAC	ACCAC
7q31	< 6.3	AC006029	CACATA	CACATA
7q31.3	< 6.3	AC02508	ATTTT	ATTTT
7q31.33	< 6.3	AC019155	TAAAG	TAAAG
9q22.2	< 6.3	AC015640	GACCC	GACCC
9q12	< 6.3	AL39220	CACTG	CACTG
9q21.12	< 6.3	AL162412	TTTTCA	TTTTCA
9q33.2	< 6.3	AL359644	TAAAAA	TAAAAA

Table B.3 ***Continued.*** Direct Repeat sequences of HERV-K(HML-2) Solitary LTRs. [a] Relative age determined by the presence or absence of the provirus in primate species (Section 3.2.2). Branching dates from humans; 55 Mya for Prosimians, 45 Mya for New World Monkeys, 28 for Old World Monkeys, 20 Mya for Gibbons, 14 Mya for Orang-utans, 8.5 Mya for Gorillas and 6.3 Mya for Chimpanzee (Figure 3.8).

HML-2 Solitary LTR	Age [a] (Mya)	Accession	5'Direct Repeat	3'Direct Repeat
9q34.13	< 6.3	AL158039	TTGGT	TTGGT
11p15.4	< 6.3	AC018539	CTATT	CTATT
11q12.3a	< 6.3	U73641	ATGTGG	ATGTGG
11q12.3b	< 6.3	AC003023	AAAC	AAAC
11q21.31	< 6.3	AP002513	TATGC	TATGC
12p11.21	< 6.3	AC068887	GGGTAC	GGGTAC
12p13.31a	< 6.3	U47924	GATATA	GATATA
12p13.31b	< 6.3	AC006432	GAGAT	GAGAT
12q13.13	< 6.3	AC027750	CTCAC	CTCAC
12q13.3b	< 6.3	AC024884	GGCACA	GGCACA
14q22.2	< 6.3	AL352982	CAAACT	CAAACT
16p12.3	< 6.3	AC002400	GTTACA	GTTACA
17p13.2	< 6.3	AC012146	TTGAC	TTGAC
17q21.2	< 6.3	AC068014	GAGAG	GAGAG
19q13.31	< 6.3	L47334	ATTTC	ATTTC
20q11.22	< 6.3	AL121753	AGAGAT	AGAGAT
21q22.3	< 6.3	Q39E10	AATCC	AATCC
Xp22.13	< 6.3	AC009858	GAAAG	GAAAG
Xq21.31	< 6.3	AL162723	GGTGGG	GGTGGG
5q23.1	< 8.5	AC008553	TTGTG	TTGTG
6q25.1	< 8.5	AC023201	ATAACA	ATAACA
7q22.3	< 8.5	AC004840	CTTTC	CTTTC
8q21.3	< 8.5	AC068510	GTGACC	GTGACC
11q13.3	< 8.5	AP002793	AAAAAC	AAAAAC
19p13.3	< 8.5	AC022148	AAAAGAG	AAAAGAG
Xp22.1	< 8.5	AC005867	GATCC	GATCC
9q34.2	< 14	AL445931	GTGGAG	GTGGAG
21q11.2	< 14	AL109748	GATTTC	GATTTC
21q21.1	< 14	AP000432	GCAGAA	GCAGAA
21q22.3	< 14	AL773587	GGTGCC	GGTGCC

5.2.1 Comparison of Direct Repeats

Table B.4 Consistent Direct Repeat sequences of HERV-K(HML-3) Proviruses.

HML-3 Provirus	Accession	5'Direct Repeat	3'Direct Repeat
1p33	AL391844	TTTTG	TTTTG
4p13	AC108467	ATTAG	ATTAG
4q13.1	AC097648	CTAAGT	CTAAGT
4q34.2	AC019163	CACAGG	CACAGG
4q35.1	AC093824	CGCAG	CGCAG
6q21	AC002464	GTGACA	GTGACA
7p13	AC073115	TATAAT	TATAAT
12q13.12	AC090058	CACCAC	CACCAC
12q23.2	AC025577	ATATAC	ATATAC
19q13.31	AC011455	CTCTGG	CTCTGG

5.2.1 Comparison of Direct Repeats

Table B.5 Consistent Direct Repeat sequences of HERV-K(HML-4) Proviruses.

HML-4 Provirus	Accession	5'Direct Repeat	3'Direct Repeat
4q13.1	AC074250	ATTCTC	ATTCTC
8q24.3	AC139103	TGCCT	TGCCT
16p13.3	AC092117	AGAGG	AGAGG
19p13.11	AC010617	GTATT	GTATT
Yq11.221	AC007034	GTATTG	GTATTG

6.2.2 and 6.2.3 Calculation of dS/dN and Maximum Parsimony

Table B.6 Accessions used to calculate dS/dN and create Maximum Parsimony Trees

HERV-K(HML-2)

HERV-K	Accession	HERV-K	Accession
HERV-K101	AC007326	HERV-K 11q22.1	AP007776
HERV-K102	AL353807	HERV-K(II)	AC092902
HERV-K103	AL591164	HERV-K 3q27.2	AC069420
HERV-K104	AC025757	HERV-K 1p31.1	AC093156
HERV-K106	AC078785	HERV-K 11q23.2	AP000831
HERV-K107	AC016577	HERV-K 10p14	AC015686
HERV-K108	AF074086	HERV-K(I)	AC084198
HERV-K109	AL590785	HERV-K 3p25	AC018829
HERV-K110	AL121985	HERV-K1 9p13.11a	AC011467
HERV-K113	AY037928	HERV-K 19q13.13	AC012309
HERV-K115	AC130464	HERV-K 6p22.1	AL121932
HERV-K 12q14.1	AC025420	HERV-K 6p21.1	AL035587

HERV-K(HML-3)

HERV-K	Accession	HERV-K	Accession
HERV-K 1p33	AL391844	HERV-K 7p13	AC073115
HERV-K 12q23.2	AC025577	HERV-K 19p13.11	AC010615
HERV-K 5q14.3	AC117524	HERV-K 4q13.1	AC097648
HERV-K 4p13	AC108467	HERV-K 12q13.12	AC090058
HERV-K 4q35.1	AC093824	HERV-K 19q13.31	AC011455
HERV-K 6q21	AC002464	HERV-K 7q21.3	AC069292

HERV-K(HML-4)

HERV-K	Accession	HERV-K	Accession
HERV-K 10p15.1	AL391427	HERV-K 8q24.3	AC139103
HERV-K 19p13.11	AC010617	HERV-K 4q13.1	AC074250
HERV-K 16p13.3	AC092117	HERV-K Yq11.22.1	AC007034
HERV-K 17q21.31	AC109326		

6.2.1 Calculation of Synonymous and Non-synonymous Distances.

Table B.7 Comparison of Methods to Calculate the Numbers of Synonymous and Non-synonymous Substitutions within HERV-K(HML-2) ORFs. SP and SP refer to the application of the Pairwise distance estimate, SJ and NJ the application of the Jukes – Cantor Model and ds and dn the Nei-Gojobori method.

Region	SP	NP	**SP/NP**	SJ	NJ	**SJ/NJ**	ds	dn	**ds/dn**
HML-2 ORFs									
ORFs (n = 24)	0.1003	0.042	**2.388**	0.1146	0.0444	**2.581**	0.112	0.044	**2.545**
Poly (n = 2)	0.0206	0.0102	**2.01**	0.0208	0.0103	**2.01**	0.02	0.01	**2**
Hsi (n = 13)	0.026	0.0137	**1.897**	0.0265	0.0139	**1.906**	0.024	0.013	**1.846**
Gorilla (n = 4)	0.2078	0.0768	**2.705**	0.2507	0.0828	**3.027**	0.239	0.073	**3.273**
Orang (n = 2)	0.0866	0.0539	**1.606**	0.092	0.0563	**1.634**	0.078	0.050	**1.56**
Gibb (n = 2)	0.1031	0.0649	**1.588**	0.1104	0.0683	**1.616**	0.099	0.059	**1.677**
HML-2 gag									
ORFs (n = 24)	0.1163	0.0531	**2.19**	0.1378	0.0573	**2.4**	0.125	0.052	**2.4**
Poly (n = 2)	0.0157	0.007	**2.242**	0.013	0.0071	**1.8**	0.016	0.006	**2.6**
Hsi (n = 13)	0.0250	0.0133	**1.879**	0.0255	0.0135	**1.8**	0.024	0.011	**2.18**
Gorilla (n = 4)	0.2585	0.1042	**2.48**	0.329	0.1157	**2.8**	0.308	0.103	**2.9**
Orang (n = 2)	0.1089	0.0584	**1.864**	0.1186	0.0615	**1.9**	0.102	0.054	**1.8**
Gibb (n = 2)	0.0975	0.0783	**1.245**	0.1048	0.0832	**1.25**	0.096	0.07	**1.3**
HML-2 prt									
ORFs (n = 24)	0.1321	0.0537	**2.459**	0.1802	0.0577	**3.123**	0.170	0.054	**3.148**
Poly (n = 2)	0.0166	0.0147	**1.129**	0.0168	0.0149	**1.127**	0.017	0.013	**1.307**
Hsi (n = 13)	0.0268	0.0184	**1.456**	0.0274	0.0187	**1.465**	0.027	0.016	**1.687**
Gorilla (n = 4)	0.3305	0.096	**3.442**	0.5072	0.1056	**4.803**	0.489	0.097	**5.041**
Orang (n = 2)	0.0947	0.0578	**1.683**	0.1017	0.0603	**1.686**	0.095	0.054	**1.759**
Gibb (n = 2)	0.1375	0.0987	**1.393**	0.1509	0.1064	**1.418**	0.142	0.095	**1.494**
HML-2 pol									
ORFs (n = 24)	0.0811	0.0309	**2.624**	0.0889	0.032	**2.778**	0.085	0.042	**2.023**
Poly (n = 2)	0.0286	0.008	**3.575**	0.0291	0.0081	**3.592**	0.025	0.008	**3.125**
Hsi (n = 13)	0.03	0.0142	**2.112**	0.0309	0.0145	**2.131**	0.028	0.013	**2.153**
Gorilla (n = 4)	0.1517	0.05	**3.034**	0.1713	0.0581	**2.948**	0.165	0.049	**3.367**
Orang (n = 2)	0.0508	0.043	**1.181**	0.0524	0.0445	**1.177**	0.021	0.044	**0.477**
Gibb (n = 2)	0.0933	0.0524	**1.78**	0.0944	0.0545	**1.732**	0.088	0.047	**1.87**
HML-2 env									
ORFs (n = 24)	0.0833	0.0385	**2.163**	0.0929	0.0406	**2.288**	0.087	0.035	**2.485**
Poly (n = 2)	0.0157	0.0137	**1.145**	0.0159	0.0138	**1.152**	0.016	0.012	**1.333**
Hsi (n = 13)	0.0266	0.0138	**1.927**	0.0271	0.014	**1.935**	0.025	0.012	**2.083**
Gorilla (n = 4)	0.1504	0.0724	**2.077**	0.1796	0.0787	**2.282**	0.166	0.069	**2.405**
Orang (n = 2)	0.0631	0.0506	**1.247**	0.0657	0.0526	**1.249**	0.057	0.045	**1.266**
Gibb (n = 2)	0.1054	0.0589	**1.789**	0.1125	0.0617	**1.823**	0.099	0.053	**1.867**

J Mol Evol (2004) 59:642–656
DOI: 10.1007/s00239-004-2656-1

JOURNAL OF MOLECULAR EVOLUTION

Allelic Variation of HERV-K(HML-2) Endogenous Retroviral Elements in Human Populations

Catriona Macfarlane, Peter Simmonds

Center for Infectious Diseases, University of Edinburgh, Summerhall, Edinburgh, Scotland EH9 1QH, UK

Received: 7 July 2003 / Accepted: 1 June 2004 [*Reviewing Editor:* Dr. Wen-Hsiung Li]

Abstract. Human endogenous retroviruses (HERVs) are the remnants of ancient germ cell infection by exogenous retroviruses and occupy up to 8% of the human genome. It has been suggested that HERV sequences have contributed to primate evolution by regulating the expression of cellular genes and mediating chromosome rearrangements. After integration ~28 million years ago, members of the HERV-K (HML-2) family have continued to amplify and recombine. To investigate the utility of HML-2 polymorphisms as markers for the study of more recent human evolution, we compiled a list of the structure and integration sites of sequences that are unique to humans and screened each insertion for polymorphism within the human genome databases. Of the total of 74 HML-2 sequences, 18 corresponded to complete or near-complete proviruses, 49 were solitary long terminal repeats (LTRs), 6 were incomplete LTRs, and 1 was a SVA retrotransposon. A number of different allelic configurations were identified including the alternation of a provirus and solitary LTR. We developed polymerase chain reaction-based assays for seven HML-2 loci and screened 109 human DNA samples from Africa, Europe, Asia, and Southeast Asia. Our results indicate that the diversity of HML-2 elements is higher in African than non-African populations, with population differentiation values ranging from 0.6 to 9.8%. These findings denote a recent expansion from Africa. We compare the phylogenetic relationships of HML-2 sequences that are unique to humans and consider whether these elements have played a role in the remodeling of the hominid genome.

Key words: Human endogenous retrovirus (HERV) — HERV-K(HML-2) — Retrovirus-like sequences — Solitary LTR — Provirus — Recombination — Gene conversion — SVA — Human genome evolution

Correspondence to: Catriona Macfarlane; *email:* catriona.macfarlane@fsamail.net

Introduction

Endogenous retroviruses (ERVs) are vertically transmitted genetic elements that remain from ancient retroviral infection of germ line cells. Following the original insertion of the provirus, intracellular retrotransposition and recombination have led to an increase in the copy number of particular families (Lower et al. 1996). ERVs are stably integrated into the genomes of all vertebrates and are transmitted as Mendelian genes. Analysis of the draft sequence of the human genome shows that approximately 8% is composed of retrovirus-like elements, which includes both proviral sequences and a large number of long terminal repeats (LTRs) (Lower et al. 1996; Patience et al. 1997; International Human Genome Sequencing Consortium 2001). Several distinct human ERV families (HERVs) have been identified, which show different genomic integration patterns (Urnovitz and Murphy 1996) and range in copy number from 1 to 1000 (Tristem 2000). HERVs are classified into families based

upon their putative tRNA primer binding site specificity; HERV-I for Ile tRNA and HERV-K for lys tRNA. Mutational events have rendered most of these HERVs replication defective following integration, although many remain transcriptionally active (Goodchild et al. 1995; Huh et al. 2003). The HERV-K superfamily is acknowledged to be the most biologically active class of HERV, having retained the ability to encode functional retroviral protein (Towler et al. 1998) and produce retrovirus-like particles (Simpson et al. 1996; Seifarth et al. 1998).

Since the identification of the HERV-K prototype, HERV-K10 (Ono et al. 1986), phylogenetic analysis of a conserved reverse transcriptase (RT) region has led to the definition of six HERV-K subgroups, HML-1 to HML-6 (Medstrand and Blomberg 1993; Zsiros et al. 1998). The HML-2 group appears to have integrated into the germ line approximately 28 million years ago, before the evolutionary split of lower Old World primates and hominoids (Reus et al. 2001b). Despite this relative age, HML-2 open reading frames appear to be maintained (Zsiros et al. 1999) and the presence of sequences that are unique to humans indicates that they were continuing to undergo amplification relatively recently (Medstrand and Mager 1998; Barbulescu et al. 1999; Buzdin et al. 2002, 2003). HML-2 proviral genomes are classified into two types based upon a 292-bp deletion at the *pol–env* boundary, with Type I elements carrying the deletion. Both Type I and Type II proviral genomes have remained retrotranspositionally active following the evolutionary split of chimpanzees and hominids (Costas 2001). HML-2 elements are easily distinguished from their progenitor, HERV-K(OLD), as they have a 96-bp deletion in *gag* which has not disrupted the open reading frame and further 8- and 23-bp deletions within their LTRs (Mayer et al. 1998; Reus et al. 2001b).

Recent genome-wide comparisons of human and chimpanzee have demonstrated that large-scale genomic rearrangements, such as segmental duplications and the insertion of retroelements, provide a significant source of DNA variation within the host species (Liu et al. 2003; Frazer et al. 2003; Locke et al. 2003). To date, most evolutionary studies have focused on the interspersed repetitive elements, L1 (long interspersed element 1) and Alu (short interspersed element); these have shown that these retroelements serve as mutagens at both the structural and genic levels (Deininger and Batzer 2002). For the same reasons, HML-2 elements may also have contributed either by serving as nucleation points for homologous recombination (Hughes and Coffin 2001) or by regulating the expression of cellular genes (Lower et al. 1996; Akopov et al. 1998; Domansky et al. 2000; Vingradova et al. 2001). In this study we have examined the genomic structure and integration sites of HML-2 elements that are unique to humans and have investigated their potential role in the remodeling of the human genome. We have also analyzed their phylogeny and demonstrated their utility for the study of human genomic diversity.

Materials and Methods

Identification of HERV-K(HML-2) Polymorphisms

The GenBank nonredundant and high-throughput genomic sequence database (http://www.ncbi.nlm.nih.gov/genome/seq/HsBlast.html), the Ensembl database (http://www.ensembl.org), and the HERV-d (http://herv.img.cas.cz) database were screened using the HERV-K10 sequence (accession No. M14123) as a probe. Accessions containing full-length HML-2 genomes were aligned by hand in the SIMMONIC sequence analysis package (Simmonds and Smith 1999), with individual elements determined by their cellular flanking sequences and chromosomal location. The flanking regions of each genome were then screened by standard nucleotide–nucleotide BLAST against the nonredundant and high-throughput sequence databases, in order to detect paralogous sequences and to ascertain if polymorphism was present at specific loci. Accessions reported to contain human-specific HML-2 LTRs and near-complete genomes were also individually screened for polymorphism within the human genome databases, with subsequent designation according to their cytogenetic location and flanking sequences.

DNA Samples and PCR Primers and Conditions

Samples from a chimpanzee (*Pan troglodytes*) and gorilla (*Gorilla gorilla*) along with 25 African, 28 Asian, 22 European, and 34 Papua New Guinean humans were collected as buccal swabs or serum. Genomic DNA was isolated using the QIAamp DNA kit (Qiagen, UK), following the manufacturer's instructions. DNA quality and authenticity were confirmed by PCR amplification for the sex chromosome-specific amelogenin gene (Faerman et al. 1995) on the human samples and the protamine gene on the chimpanzee samples (data not shown).

Each sample was subjected to a series of PCR amplification reactions in order to assess polymorphism within selected HML-2 loci (Fig. 1). DNA sequences adjacent to each HML-2 insertion were used to design unique flanking region primers. Primers were screened by standard nucleotide–nucleotide BLAST against the nonredundant and high-throughput sequence databases, to ensure that the DNA sequences were unique. Elements that resided in repetitive sequence regions could not be examined by PCR. Universal primers for HML-2 LTR, *gag*, and *env* genes were designed according to a consensus sequence, which was obtained by aligning all of the HML-2 sequences examined in this study. Heminested PCR reactions were performed in instances where single-round PCR proved difficult to optimize. This process utilized two consecutive rounds of amplification, the first round using an external pair of primers while the second round contained one of the first primers and a single nested primer which is internal to the first primer pair. The amplicon produced by the first round of PCR was used as a template for the second PCR amplification.

PCR amplification primers and conditions for each HML-2 loci are listed in Table 1. Reactions were carried out in volumes of 50 µl, with each containing 200 ng of genomic DNA, a 200 µ*M* concentration of each dNTP, a 0.5 µ*M* concentration of each primer, and 0.5 units of *Taq* DNA polymerase in standard PCR buffer as supplied by Promega. The second round of PCR used 2 µl of first-round PCR product and was performed in volumes of 30 µl, with a reaction mix as listed above. The resulting PCR products were analyzed

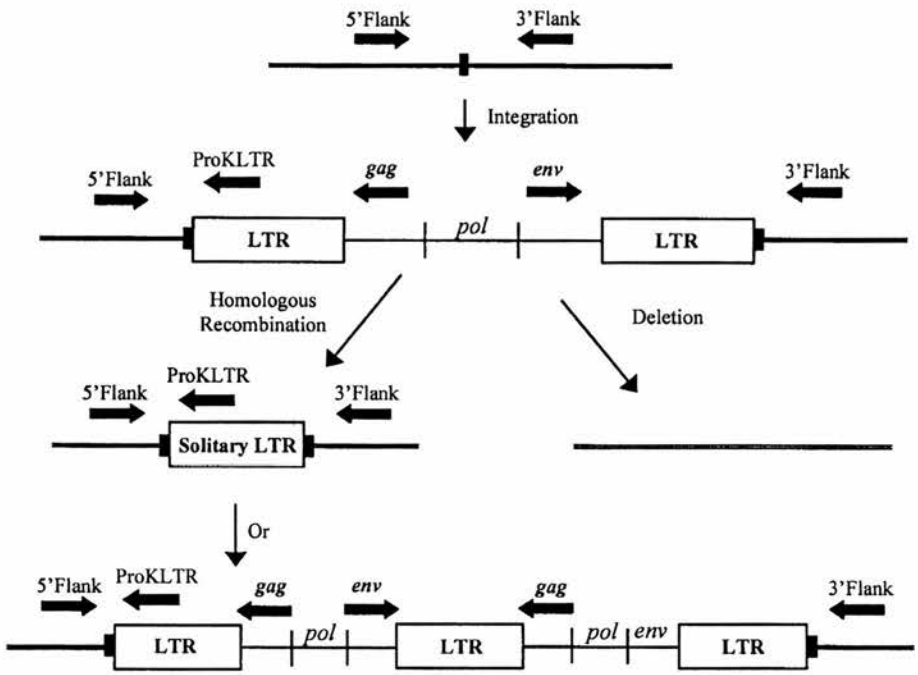


Fig. 1. PCR scheme for detecting HERV-K(HML-2) allelic variants.

by electrophoresis through a 2% agarose gel, with the product size confirmed by comparison to a 100-bp ladder (Promega). Nucleotide sequencing was carried out directly on second-round PCR products using ABI PRISM Big Dye kits (Applied Biosystems).

Sequence and Population Genetic Analysis

Sequence data obtained using the ABI PRISM kits were viewed using the CHROMAS sequence viewer and directly imported into the SIMMONICS sequence analysis package. Eighty-six full-length LTRs which were representative of 67 human-specific HML-2 insertions were aligned by hand in SIMMONIC. A neighbor-joining tree was constructed using MEGA, version 2.1 (http://www.megasoftware.net/), with the Kimura two-parameter distance estimate and pairwise deletion of gaps. Allele frequencies, Hardy–Weinberg tests, and Wright's F_{st} statistic were estimated using PopGene, version 1.31 (http://www.ualberta.ca/~fyeh/).

Results

Activity of HERV–K(HML-2) Elements Within the Human Lineage

Screening of the human sequence databases for HML-2 proviruses revealed 3 novel sequences and 29 formerly identified complete and near-complete elements (Table 2). A further 15 less intact proviral sequences have also previously been identified (Hughes and Coffin 2001; Reus et al. 2001b), bringing the total number of identified HML-2 proviruses within the human genome to 47. The three HML-2 near-complete proviruses identified in this study were located at 4p16 (AC105916), Xq28 (AF277315), and 10q24.2 (AL392107). The chimpanzee orthologue of the provirus contained within the pseudoautosomal region of the human X chromosome (Xq28) was also detected within accession AC144385.

Eight of 18 human-specific HML-2 proviral genomes were Type I and 10 were Type II, indicating the coexistence of two retrotranspositionally active master elements during hominid speciation. Further computational screening with the flanking regions of individual elements revealed that five were polymorphic, showing a number of different configurations (Fig. 1). Two proviruses, HERV-K113 and HERV-K115, were dimorphic for insertion, with one allele representing the presence and the second the absence of a complete provirus. Other variable features included the alternation of a complete provirus with a solitary LTR (HERV-K103 and HERV-K106) and, finally, the variable existence of a tandem duplication of the HERV-K108 provirus.

Human-specific HML-2 LTR sequences have previously been identified by targeted genomic difference analysis (TGDA) and BLAST determination, with subsequent phylogenetic identification by PCR amplification (Buzdin et al. 2002; Lebedev et al. 2000). We catalogued all of the human-specific HML-2 LTR sequences discovered to date, determined their cytogenetic location, and assigned subtype according to their classification in previous publications (Table 3). During sequence alignment we observed that several LTR sequences either had been observed by more than one study and were assigned different names or were misinterpreted as solitary LTRs when they were part of a complete provirus. Of the total of 74 human-specific HML-2 LTR sequences, 18 were complete or near-complete

Table 1. PCR primers and annealing temperatures

PCR amplification	5′ primer sequence	3′ primer sequence	AT[a]
K113 insertion site	TGCATGGGGAGATTCAGAACC	ATCCATACATTTCTGAGTCCTGA	56
K113 LTR	TGCATGGGGAGATTCAGAACC	AATGGAGTCTCCYATGTCTACT	56
K113 full provirus	TGCATGGGGAGATTCAGAACC	GGATCTCTYGTCGACTTGTCC	58
K115 insertion site	AGCACTGAGATCCAAACTCATAT	CAGTCTATAGATGTGGATGCCT	58
K115 LTR	AGCACTGAGATCCAAACTCATAT	AGGGMGTRGTGATGACTCTTAA	58
K115 full provirus	AGCACTGAGATCCAAACTCATAT	GGATCTCTYGTCGACTTGTCC	58
K103 insertion site	CCACCATCTGAGAAGTGTGATG	GGCAACAAAGGGTTCATATGAGAA	50
K103 LTR	CCACCATCTGAGAAGTGTGATG	AATGGAGTCTCCYATGTCTACT	50
K103 full provirus	CCACCATCTGAGAAGTGTGATG	GGATCTCTYGTCGACTTGTCC	58
K103 solitary LTR	CCACCATCTGAGAAGTGTGATG	GGCAACAAAGGGTTCATATGAGAA	58
K106 insertion site	TCCACCTGCGGACCTCCTCT	TATTGGTGACAGAGAGATGCAG	58
K106 LTR	TCCACCTGCGGACCTCCTCT	AATGGAGTCTCCYATGTCTACT	58
K106 full provirus	TCCACCTGCGGACCTCCTCTA TTCCACCAGCCTGTAGGGGA	GGATCTCTYGTCGACTTGTCC	58
K106 solitary LTR	TCCACCTGCGGACCTCCTCTATT CCACCAGCCTGTAGGGGA	TATTGGTGACAGAGAGATGCAG	58
K107 insertion site	GGACACCCAACCTGCATGGT	ACACCACTGACAGTTACAGTACC	58
K107 LTR	GGACACCCAACCTGCATGGT	AATGGAGTCTCCYATGTCTACT	58
K107 full provirus	GGACACCCAACCTGCATGGTTC AACTCACTGCTGTGGGGAA	GGATCTCTYGTCGACTTGTCC	58
K107 solitary LTR	GGACACCCAACCTGCATGGTTC AACTCACTGCTGTGGGGAA	GCCGGAGGTTGTGTAGGGG	58
K108 LTR	GTTACAGGAGTGCGCCATCAC	AGGGMGTRGTGATGACTCTTAA	58
K108 full provirus	GTTACAGGAGTGCGCCATCAC	GGATCTCTYGTCGACTTGTCC	58
K108 tandem repeat	GGATCTCTYGTCGACTTGTCC	GCAGGTKTAMCCAACAGCTC	58
K108 solitary LTR	GTTACAGGAGTGCGCCATCACA GAGATGGGTTTCTGTGGGGA	GAATTAGGCTTTCGGGACTT CAGATGGTGGAAACCTGTAGGGGG	58
3q27.2 LTR	TGAGACAGGTACATGTGGGGAA	AGGGMGTRGTGATGACTCTTAA	58
3q27.2 full provirus	TGAGACAGGTACATGTGGGGAA	GGATCTCTYGTCGACTTGTCC	58
3q27.2 solitary LTR	TGAGACAGGTACATGTGGGGAA	GTATTTTATGTTATGTACCTGTAGG	58
7p21.2 insertion site	CCACTGTGTACAAGTATATGTG GAGTCAGGGTCTCTTCTGTTG	GATTGCTCTTATAAGTCAGTTTGA	50
7p21.2 LTR	CCACTGTGTACAAGTATATG TGGAGTCAGGGTCTCTTCTGTTG	AATGGAGTCTCCYATGTCTACT	50
17q22 insertion site	GATTGCTCTTATAAGTCA GTTTGAGGGATCTTACAGATACACCAGT	GGGTGCAGCACACCAACATG	50
17q22 LTR	GATTGCTCTTATAAGTCAGTTTGAGGGA TCTTACAGATACACCAGT	AATGGAGTCTCCYATGTCTACT	50

[a]Amplification required 2 min of initial denaturing at 94°C, and 35 cycles of 30 s at 94°C, 30 s at the annealing temperature (AT), and 30 s of elongation at 72°C. A final extension time of 6 min at 72°C was added.

proviruses, 49 were solitary LTRs, and a further 6 could not be distinguished between near-complete proviral sequences and solitary LTRs, as they have lost the 5′ or 3′ end of their sequence. Further sequence comparison of the HML-2 LTR contained at Xq26.3 (AL359703) to the SVA_{STPA1} retroelement (AC016142) (Ostertag et al. 2003) revealed that this human-specific sequence was a member of the SVA (SINE, VNTR, and Alu) retrotransposon family. As SVA elements are derived from SINE.R retroelements which are composed of a partial HERV-K(HML-2) *env* and a 3′-LTR (Shen et al. 1994), it can be concluded that the LTR at Xq26.3 is not a direct product of the retrotransposition of a HERV-K (HML-2) provirus. Computational screening of the flanking regions of each of the human-specific HML-2 LTRs indicated that two solitary LTRs were polymorphic for insertion. The first was located at 6p21.32 (Z80898) and is reported to have arisen through duplication of the MHC complex (Horton et al. 1998); the second was located at 9q12 (AL39220). With the exception of chromosomes 13, 15, 18, and Y, all chromosomes contained at least one human-specific HERV-K(HML-2) LTR sequence that arose through the process of retrotransposition.

Table 2. Complete and near-complete HERV-K(HML-2) proviruses within the human genome

HERV	Species[a]	Location	Type[b]	Accession No.	Nucleotide difference	Features[c,d]	Reference
K101	Human	22qll.2	I	AF16409	2		Barbulescu et al. (1999)
				AC007326/FID 83799	5		
K102	Human	1q21	I	AFI64610	4		Barbulescu et al. (1999)
				AL353807/FID 1	2		
				AC044819	2		
K103	Human	10p12.1	I	AF164611/AF59796	7		Barbulescu et al. (1999)
				AL591164	6		
				AL139404	Solo LTR	Polymorphic	This study
K104	Human	5p14.3	II	AF164612	17		Barbulescu et al. (1999)
				AC025757/AC116309	17		
K106	Human	3ql3.2	I	AF16540/AC078785	1		Barbulescu et al. (1999)
				AC024108	Solo LTR	Polymorphic	This study
K107	Human	5q33.3	I	M14123	2		Ono (1986)
HERV-K10				AF164613	4		
				AC016577/FID27409	2		
K108	Human	7p22.1	II	AC072054/AC0104060	2	Polymorphic	Mayer et al. (1999)
HML-2.HOM				Y17832/AF164614	6		Tonjes et al. (1999)
HERV-K(C7)				AF074086	0		Reus et al. (2001a)
				FID37994	3		
				AF261945			
K109	Human	6q14.1	II	AL590785	2		Barbulescu et al. (1999)
				AC0055116	5		
K113	Human	19p13.11	II	AY037928	3	Polymorphic	Turner et al. (2001)
				AC092364			
K115	Human	8p23.1	II	AY037929	14	Polymorphic	Turner et al. (2001)
				AC130464/AC130367	14		
12q14.1	Human	12q14.1	II	AC0025420	4		Costas (2001)
				AC074261	19		
				FID58908	20		
11q22.1	Human	11q22.1	II	AP007776/FID54721	4		Costas (2001)
HERV-K(II)	Human	3q21.2	II	AB047209/AC092902	18		Sugimoto et al. (2001)
				AC069047/AC092903	18		
				AC026957			
3q27.2	Human	3q27.2	I	AC069420	3		Hughes and Coffin (2001)
				AC015525/AC133473			
1p31.1	Human	1p31.1	I	AC093156	0	Δ2846 bp *pol*	Hughes and Coffin (2001)
21q21.1	Human	21q21.1	I	AL109763		Δ164 bp gag	Kurdyukov et al. (2001)
				AL163218		Δ712 bp 3′ LTR	
				AF240627			
HERV-K(C19)	Human	19p12-q12	II	AFO17229		Δ5′ LTR	Tonjes et al. (1999)
				AC112702/			
				AC010508			
				Y 17833			
12q24.11	Human	12q24.11	II	AC002350		Δ520 bp *env*	Medstrand and Mager (1998)
						3′ LTR	
4q23.1	Chimp	4q32.1	I	AC106872	37	Δ1937 bp *pol*	Hughes and Coffin (2001)
				AC108519/AC068369			
K105	Gorilla	21q11.1	I	AF16419	40		Barbulescu et al. (1999)
				AF260249			
				AF260253			
K110	Gorilla	1q23.3	I	AL121985/AC068728	34		Ono (1986)
HERV-K18				Y18890/FID2	33		
				AF164618	36		
				AF134984/AF012336			
11q23.2	Gorilla	11q23.2	I	AP000831/FID54716	6		Costas (2001)
10p14	Gorilla	10pl4	II	AC015686/FID50753	30		Costas (2001)
				AL392086	30		
HERV-K(I)	Gorilla	3q12.1	I	AB047240	19		Sugimoto et al. (2001)
				AC084198/FID13837	18		
6p21.1	Gorilla	6p21.1	II	AL035587	39	Ins Alu Y *gag*	Reus et al. (2001b)
3p25	Orangutan	3p25	I	AC018829/AC018809	53	Δ2554 bp *pol*	Hughes and Coffin (2001)
19p13.11	Orangutan	19p13.11	I	AC011467/AC036240	62	Ins 6760 bp	Hughes and Coffin (2001)
				AC068369/AC078899		5′ LTR Δ2554 bp *pol*	

(Continued)

Table 2. Continued

HERV	Species[a]	Location	Type[b]	Accession No.	Nucleotide difference	Features[c,d]	Reference
19q13.13	Gibbon	19ql3.13	II	ACO12309	78		Reus et al. (2001b)
6p22.1	Gibbon	6p22.1	II	AL121932 AL390196/ AL671879	60	Ins solo LTR *pol*	Reus et al. (2001b)
4p16		4p16	II	AC105916	70		This study
Xq28		Xp28	II	AF277315	49	Δ2181 bp *gag–pol*	This study
10q24.2		10q24.2	II	AL392107	33 (335)	Ins 9598 bp 5′ LTR Δ634 bp 5′LTR Δ1375 bp *env*	This study

[a]The most distant species in which a given provirus is reported to be found (Barbulescu et al. 1999; Kurdyukov et al. 2001; Hughes and Coffin 2001; Turner et al. 2001).
[b]HERV-K(HML-2) proviral genome classification, based upon the presence of the diagnostic, 292-bp deletion at the *pol–env* boundary, with Type I elements carrying the deletion.
[c]Ins, insertion.
[d]Δ, deletion.

Inter- and Intraelement Recombination—Comparison of Direct Repeat Sequences

In common with infection by exogenous retroviruses, retrotransposition and subsequent integration of HERV-K endogenous retrovirus result in the generation of short target site duplications of 4 to 6 bp at the integration site. These direct repeat sequences flank either end of the newly integrated provirus and should be identical at the time of integration. Integrated proviruses or solitary LTRs with different target site duplications serve as potential signatures of interelement homologous recombination, which would be expected to have resulted in large-scale chromosomal rearrangements (Hughes and Coffin 2001). We examined the target site duplications of the 74 human-specific HML-2 sequences in order to investigate the impact of such events during hominid evolution. The three near-complete proviruses, 12q24.11 (AC002350); HERV-K(C19) (AF017229), and 21q21.1 (AL109763), along with the six incomplete LTRs, were not included in the data set, as they had lost one or more of their direct repeats. The human-specific HML-2 LTR at Xq26.3 (AL359703), which was a component of a SVA retroelement, and the polymorphic solitary LTR at 6p21.32 (Z80898) were also not included, as they did not represent the recent retrotranspostion and integration of a proviral sequence. This left a total of 63 HML-2 elements that could be analyzed.

Each of the respective allelic variants of the polymorphic loci, HERV-K103 and HERV-K106, had identical direct repeats, implying that the solitary LTRs were generated as a result of (intraelement) homologous recombination between the 5′- and the 3′-LTRs of each respective provirus. Such an event is also expected to produce and result in the loss of a near-complete provirus consisting of a single LTR along with the *gag*, *pol*, and *env* genes. For the same reasons, intraelement recombination leading to the internal duplication of a proviral sequence is also likely to have generated the tandem duplication of the HERV-K108 provirus and is expected to have also led to the formation of a solitary LTR at the same chromosomal location.

If HERV-K sequences with inconsistent direct repeats arose through (interelement) homologous recombination between proviruses located either on different chromosomes or in different regions of the same chromosome, then exchanges would be expected to produce a reciprocal HERV-K element with an opposite configuration of direct repeats and flanking regions. Of the remaining HML-2 sequences, two had disparate target site duplications, indicating their likely hybrid nature. Within the human genome databases they exist as solitary LTRs and are located on chromosomes 7p21.2 (AC006035) and 17q22 (AC032016).

We screened the human genome databases for the expected reciprocal product of each of the "hybrid" HML-2 sequences; none of the predicted sequences were present. This implied either that the reciprocal products were not present in the representative individuals sequenced by the human genome projects or that the expected reciprocal sequences do not form a constituent of the contemporary human gene pool. In order to confirm that the two human-specific solitary LTRs were a product of interelement recombination, we designed unique 5′ and 3′ flanking region primers for the solitary LTRs at 7p21.2 (AC006035) and 17q22 (AC032016) and conducted amplification for both the solitary LTR and the preintegration site in human and

Table 3. HERV-K(HML-2) LTR sequences that are unique to humans

Location	Accession no.	Features	Subfamily[a]	Reference
1p22.1	AF370125/AL139421		LTR II-L/HS-a	Buzdin et al. (2002)
1p31.2	AL356736/AL359701		LTR II-L/HS-a	Buzdin et al. (2002)
1q22	AL135927			Buzdin et al. (2003)
2p22.2	AC007390			Buzdin et al. (2003)
2p23.14	AC021294		LTR II-L/HS-a	Buzdin et al. (2002)
2p23.3	AC074117		LTR II-L/HS-a	Buzdin et al. (2002)
2q21.2	AC084028/AC093787	ΔLTR[c]	LTR II-L/HS-a	Buzdin et al. (2002)
2q33.2	AC074019		LTR II-T	Mamedov et al. (2002)
3p12.3	AF042089		LTR II-L	Buzdin et al. (2002)
3p21.31a	Z84493/AL450422		HS-a	Medstrand and Mager (1998)
3p21.31b	AC025548/AC104447			Buzdin et al. (2003)
3q26.31	AC068566/AC104640		LTR II-L/HS-b	Buzdin et al. (2002)
3q28	AC0620087/AC112909		LTR II-L	Buzdin et al. (2002)
4q13.3	AC055844/AC106051			Buzdin et al. (2003)
5p15.31	AC091985		LTR II-L4	Mamedov et al. (2002)
5q23.1	AC010267		LTR II-L/HS-b	Buzdin et al. (2002)
5q35.1	AC008648			Buzdin et al. (2003)
5q35.3	AC023559/AC113425		LTR II-L/HS-a	Buzdin et al. (2002)
6q15	AL021774/AL139090		LTR II-L/HS-a	Buzdin et al. (2002)
6q23.2	AL596188		LTR II-L4	Mamedov et al. (2002)
6p21.32a	Z80898/U92032	Polymorphic	HS-b	Horton et al. (1998)
6p21.32b	AC022567/X87344		LTR II-T	Buzdin et al. (2002)
7p21.2	AC006035	Direct repeats vary	LTR II-L4	Mamedov et al. (2002)
7q31	AC006029		LTR II-L/HS-a	Buzdin et al. (2002)
7q31.3	AC02508		LTR II-L3/HS-a	Medstrand and Mager (1998)
7q31.33	AC019155		LTR II-L4	Mamedov et al. (2002)
9q22.2	AC015640/AL590377			Buzdin et al. (2003)
9q12	AL39220/AL773545	Polymorphic[b]	LTR II-L/HS-a	Buzdin et al. (2002)
9q21.12	AL162412		LTR II-L/HS-a	Buzdin et al. (2002)
9q33.2	AL359644		LTR II-L4	Mamedov et al. (2002)
9q34.13	AL158039/AL354855		LTR II-L/HS-a	Buzdin et al. (2002)
11p15.4	AC018539/AC080023		LTR II-L4	Mamedov et al. (2002)
11q12.3a	U73641/AP001591		HS-a	Medstrand and Mager (1998)
11q12.3b	AC003023/AP002793		LTR II-L	Buzdin et al. (2002)
11q13.3	AP001184	ΔLTR[c]	LTR II-L/HS-a	Buzdin et al. (2002)
11q21.31	AP002513/AC021821		LTR II-L4	Mamedov et al. (2002)
12p11.21	AC068887/AC048344			Buzdin et al. (2003).
12p13.31a	U47924		HS-b	Medstrand and Mager (1998)
12p13.31b	AC006432			Buzdin et al. (2003)
12q13.13	AC027750/AC107031			Buzdin et al. (2003)
12q13.3	AC079034	ΔLTR[c]	LTR II-L/HS-a	Buzdin et al. (2002)
12q13.3	AC024884/AC025574		LTR II-L/HS-a	Buzdin et al. (2002)
14q22.2	AL352982			Buzdin et al. (2003)
14q23.3	AL139022	ΔLTR[c]	LTR II-L/HS-a	Buzdin et al. (2002)
16p12.3	AC002400/AC008870		HS-b	Medstrand and Mager (1998)
16p13.12	AC009167	ΔLTR[c]	LTR II-L	Buzdin et al. (2002)
16q23.1	AC009132	ΔLTR[c]	LTR II-L4	Mamedov et al. (2002)
17p13.2	AC012146		LTR II-L/HS-b	Buzdin et al. (2002)
17q21.2	AC068014		LTR II-L	Buzdin et al. (2002)
17q22	AC032016/AC000389	Direct repeats vary	LTR II-L/HS-b	Buzdin et al. (2002)
19q13.31	L47334/AC073898		LTR II-L2/HS-b	Medstrand and Mager (1998)
20q11.22	AL121753			Buzdin et al. (2003)
21q22.3	Q39E10/AF260248		LTR-L/HS-a	Kurdyukov et al. (2001)
Xp22.13	AC009858/AL732371		LTR II-L	Buzdin et al. (2002)
Xq21.31	AL162723		LTR II-L2	Mamedov et al. (2002)
Xq26.3	AL359703	3′LTR (SVA)	LTR II-L	Buzdin et al. (2002)

[a]Subfamily classification as defined (Lebedev et al. 2002; Mamedov et al. 2002; Buzdin 2003).
[b]This study.
[c]Deletion.

nonhuman primates. Initial results confirmed that both of the solitary LTRs were not present in chimpanzee and gorilla, indicating either that they were not fixed in the gene pool at the time of human/chimpanzee/gorilla divergence or that they had integrated during hominid evolution. We then performed amplification for the

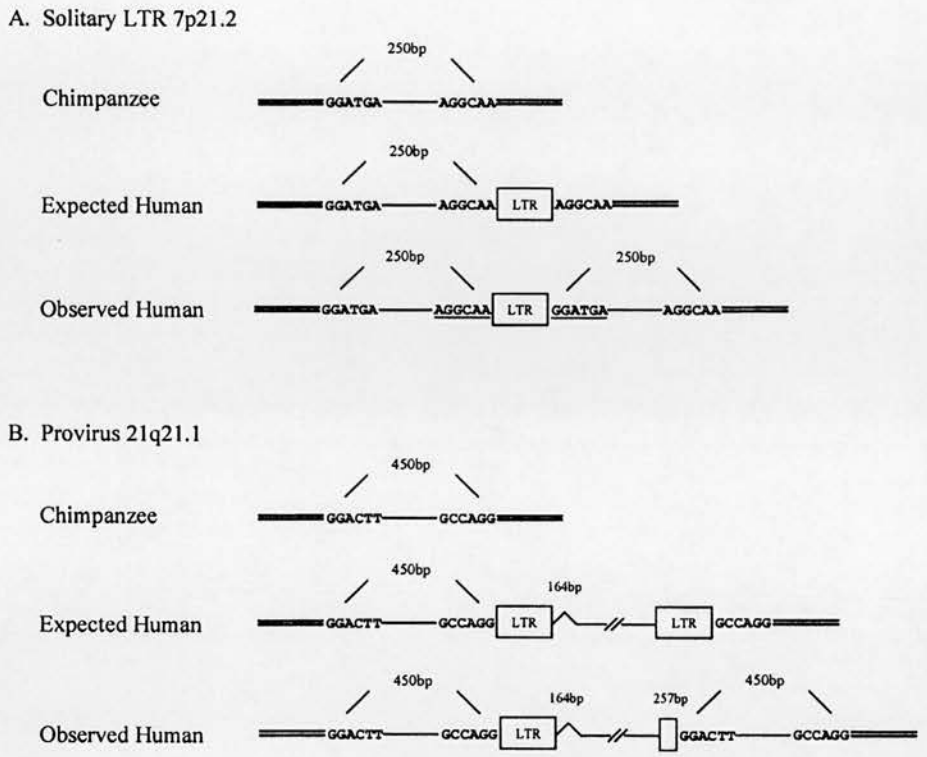


Fig. 2. Flanking region duplication leading to variable direct repeat sequences. **A** Solitary LTR 7p21.2. The preintegration sequence in chimpanzee and gorilla is represented by the top figure and contains a 250-bp sequence with the respective nucleotide sequences GGATGA and AGGCAA at each end. The human-specific solitary LTR at 7p21.2 (accession No. AC006035) has inconsistent direct repeat sequences AGGCAA and GGATGA, which are underlined. Sequence comparison indicates that the solitary LTR is flanked by a 250-bp duplication of the preintegration sequence. **B** Provirus 21q21.1. The top figure represents the preintegration sequence present in chimpanzee (accession No. BS000043). The near-complete human-specific provirus at 21p21.1 (accession No. AL109763) contains a truncated 3′-LTR of 257 bp which is adjacent to a 450-bp duplication of the preintegration site sequence.

```
Human      ATATCTACAAGGTTATTAATTGCAACACTTTTTATAATAACAAAT LTR 17q22 ACGATTTAGAATATCTTTTTTGTATTATACTTTAAGTTTTA
Chimpanzee .........G.....------------------.....
Gorilla    .........G.....------------------.....
```

Fig. 3. Sequence alignment of the human-specific HERV-K (HML-2) solitary LTR at chromosomal location 17q22 and preintegration site in primates. The human sequence of the HERV-K(HML-2) LTR-containing locus (accession No. AC032016) is shown on the top line. Nucleotide substutions at each position are indicated with the appropriate nucleotide. Alignment gaps are indicated by dashes. The direct repeats of the solitary LTR are underlined.

preintegration site in the nonhuman primates under the expectation that a negative result indicated interelement recombination. If recombination had occurred between different proviruses located at different chromosomal regions, then the expected product either would be too large to amplify or would not exist in nonhuman primates. Contrary to expectation, amplicons were produced, suggesting that the disparate target site duplications were generated by an alternative mechanism (Figs. 2 and 3).

Sequence analysis of the preintegration site and solitary LTR at 7p21.2 indicated that the variable direct repeats were a result of an apparent duplication of the 5′ flanking sequence (Fig. 2A). A similar situation was also observed for the human-specific provirus at 21q21.1, where the 3′-LTR of the provirus appears to be truncated by a sequence paralogous to the 5′ flanking sequence (Fig. 2B). Presuming that the provirus contained two identical LTRs at insertion, this duplication must have occurred following integration. For the solitary LTR at 17q22, sequence data on the preintegration site indicated that the downstream direct repeat was 4 bp shorter than the upstream one (Fig. 3). These observations indicate that inconsistent direct repeat sequences do not always reflect interelement recombination events (see Discussion).

Phylogeny of HML-2 LTRs Unique to Humans

In order to further examine the retrotransposition and evolution of the human-specific HML-2 elements, we generated a neighbor-joining tree from the alignment of 87 full-length HML-2 LTR sequences (Fig. 4). Individual LTR sequences are identified according to their consensus name or genomic location. With the exception of the HERV-K115 provirus, the 5′- and 3′-LTR of each individual HML-2 provirus grouped together, supporting the view that they had not undergone interelement recombination or sequence exchange. However, the LTRs of the HERV-K115 provirus were divergent, reflective of the provirus having undergone gene conversion. The solitary LTRs of the polymorphic loci HERV-K103 and HERV-K106 grouped with the LTRs of their progenitor provirus, confirming that they were generated through intraelement homologous recombination. The respective 5′- and 3′-LTRs of the

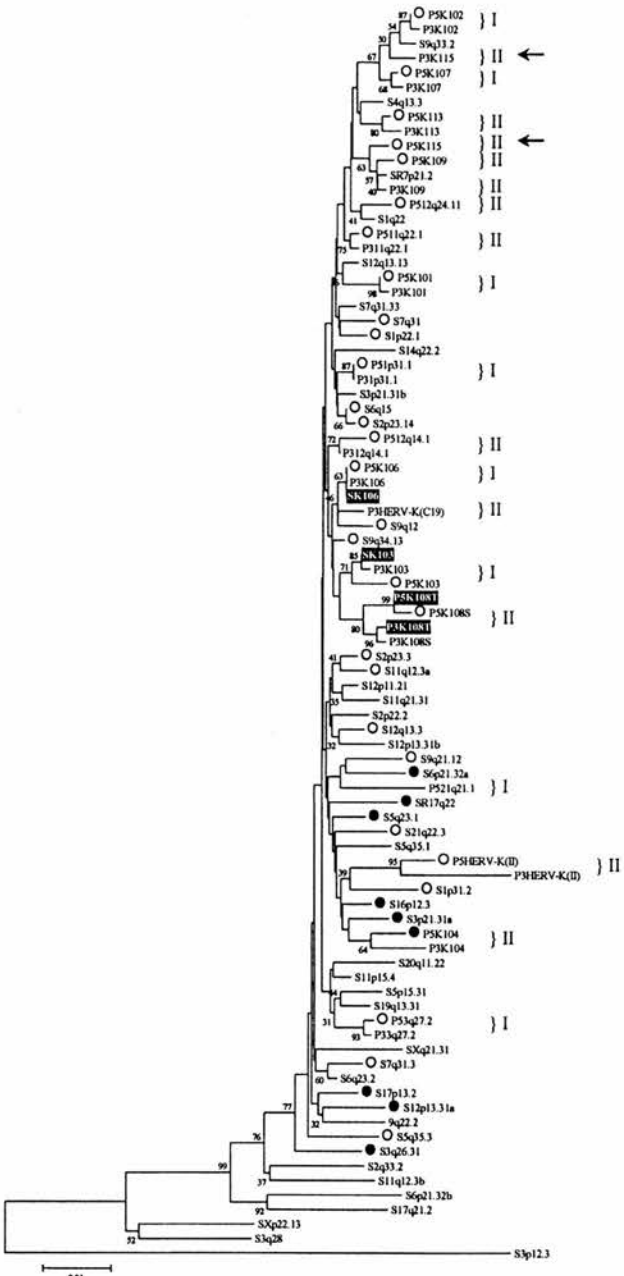


Fig. 4. Phylogeny of human-specific HML-2 LTRs. Individual LTRs are named according to the chromosomal location of the corresponding accession clone or bibliographic name of the sequence. P5—5-LTR; P3—3-LTR; S—Solitary LTR; R—direct repeats vary. The shaded LTRs are polymorphic and have arisen through intraelement recombination. The arrows emphasize the divergent LTRs of the HERV-K115 provirus. Roman numerals denote the genome structure of the HERV-K(HML-2) proviral sequences, with Type I sequences carrying a 292-bp deletion at the *pol–env* boundary. (○) LTR subfamily HS-a. (●) LTR subfamily HS-b.

tandemly repeated provirus and the single, provirus located at 7p22.1 (HERV-K108) also grouped, indicating that the tandem proviral sequence also arose through intraelement recombination. Interestingly, the distribution of the Type I and Type II proviral LTRs was not monophyletic, as would be expected if these elements follow a "master" or "source" model of retrotransposition (see Discussion).

HERV-K LTRs have previously been classified according to diagnostic nucleotide differences and intragroup divergence. Recently, 40 human-specific HML-2 LTR sequences were classified into the subtypes Hs-a and Hs-b (Buzdin et al. 2003). Superimposition of these subtypes on the phylogenetic tree indicated that this taxomony was consistent, although the two subtypes were not clearly distinguished and grouped independently of the Type I and Type II proviral genomes. This suggests that HML-2 LTRs may be subject to a high degree of sequence exchange between closely related sequences.

Relative Age of HML-2 Loci

During the retrotransposition of a provirus, reverse transcription generates a new retrovirus-like sequence containing two identical LTR sequences. Assuming that a provirus has not undergone any form of sequence exchange and there is no selective pressure acting on it, the accumulative nucleotide differences between the LTRs can serve as a molecular clock (Dangel et al. 1995). We calculated the number of nucleotide differences between the 5′- and the 3′-LTRs of the 32 intact HML-2 proviruses that were present within the human genome and compared each result to the relative age of the provirus (Table 2). Several inconsistencies were observed which were indicative of sequence homogenization between the LTRs of individual elements.

First, sequence data for individual elements were discrepant; this is exemplified by the human-specific provirus at 12q14.1 where the accession clone AC025420 has 4 nucleotide differences but the analogous accessions AC074261 and FID58908 had 19 and 20 differences, respectively. Second, relative age did not correspond with accumulative nucleotide differences. As the provirus at 11q23.2 (AP000831) is present in humans, chimpanzees, and gorillas, its relative age is expected to be between 6.2 and 12 mya (Chen and Li 2001). However, it contains only six accumulative nucleotide differences between the LTRs, which, compared to the relative age of comparable proviral loci such as 10p14 (AC015686), is indicative of a more recent insertion. Additionally, the LTRs of the insertionally polymorphic HERV-K113 provirus (AY037928) vary by 3 bp, whereas the HERV-K106 provirus (AC078785), which is reported to be universal in contemporary humans (Barbulescu et al. 1999), has only a 1-bp difference between the LTRs. In order to further investigate these discrepancies we designed PCR-based assays and tested for the absence of two human-specific proviral loci; 3q27.2 (AC069420) and HERV-K107 (5q33.3), which, according to their accumulative differences

Table 4. Genomic variation associated with the HERV-K108 locus

Population	*n*	AA	AB/BB	AC
Africa	25	8	16	1
Asia	28	17	11	0
Europe	21	7	14	0
Papua New Guinea	17	12	5	0
Total	91	44	46	1

Note. A, single provirus; B, tandemly repeated provirus; C, solitary LTR.

Table 5. Genomic variation associated with HERV-K(HML-2) loci

	HERV-K locus											
	K113[a]			K115[a]			K103[b]			K106[b]		
Population	*n*	(+) Frequency	*h*[c]	*n*	(+) Frequency	*h*	*n*	(+) Frequency	*h*	*n*	(+) Frequency	*h*
Africa	25	0.2	0.32	25	0.2	0.32	25	0.04	0.08	25	0.1	0.18
Asia	28	0.107	0.19	28	0.0	0.0	28	0.0	0.0	27	0.074	0.1
Europe	22	0.0	0.0	22	0.0	0.0	22	0.0	0.0	21	0.071	0.09
Papua New Guinea	26	0.231	0.36	34	0.088	0.18	15	0.0	0.0	28	0.071	0.13
Average		0.134	0.217		0.072	0.125		0.01	0.02		0.079	0.125
Total	101	0.138	0.216	109	0.073	0.126	90	0.011	0.019	101	0.069	0.126
F_{st}		0.069			0.098			0.03			0.006	

[a]Provirus insertion.
[b]Solitary LTR.
[c]Heterozygosity.

(three and two), appear to have integrated into the human germ line relatively recently.

PCR Analysis of HML-2 Loci

HERV-K polymorphisms serve as ideal genetic markers for examining human evolution, as they are stable and identical by descent and the ancestral state is known to be the absence of the insertion. We developed a PCR-based assay to examine the allelic variation associated with seven HML-2 proviral loci—HERV-K113 (19p12), HERV-K115 (8p23.1), HERV-K103 (10p12.1), HERV-K106 (3q13.3), HERV-K108 (7p22.1), 3q27.2 (AC069420), and HERV-K107 (5q33.3)—and determined their geographical distribution by amplifying for their presence in 109 DNA samples from four diverse human populations. A schematic diagram of our PCR-based strategy and the predicted outcomes of intraelement homologous recombination are depicted in Fig. 1. Unique 5′ and 3′ flanking region primers were designed in order to detect the preintegration site sequence or solitary LTR at each of the selected loci. The absence of PCR product indicated either the deletion or the presence of a HML-2 provirus. This was evaluated by amplifying for a complete proviral sequence using the unique 5′ flanking primer and universal *gag* primer. Detection of the allelic variation present at the HERV-K108 loci, which can contain a tandemly repeated provirus (Reus et al. 2001a), initially involved conformational screening for the presence of proviral sequence at that locus, using the unique 5′ flanking region primer and universal *gag* primer. The presence of at least one copy of the tandemly repeated provirus was analyzed by amplifying with the universal primers *gag* and *env*. Computational screening within the human genome databases for the potential combinations of the universal primers *gag* and *env* indicated that the predicted amplicon was unique to the HERV-K108 on chromosome 7.

The first polymorphic HML-2 locus that we examined was HERV-K108 on chromosome 7. As we did not perform amplification reactions that spanned the entire length of the HERV-K108 loci, we were unable to distinguish between individuals who were heterozygous in possessing one copy of the ancestral single proviral allele (A) and a copy of the tandemly repeated provirus (B) from individuals who were homozygous for the tandemly repeated provirus (BB) (Table 4). However, in performing conformational amplification for the presence of the HERV-K108 insertion, we were able to determine the number of individuals who were homozygous in possessing the ancestral copy of the provirus (AA). Interelement recombination leading to the production of a tandemly repeated provirus is also expected to generate a solitary LTR. We detected such a solitary LTR in a single individual, indicative of an allele frequency of

0.02 within the African population and a worldwide frequency of 0.005.

The human genomic variation associated with the remaining six HML-2 loci indicated that four of the loci were dimorphic and two loci were monomorphic, consistent with the data retrieved from the human genome databases (Table 5). Allele frequencies for the bi-allelic loci ranged from 0.231 for the HERV-K113 provirus in the Papua New Guinean population to 0.00 for all loci in a number of cases. Interestingly, the allele frequencies for the solitary LTR at the HERV-K103 locus ranged from 0.04 in the African population to zero in all other populations, perhaps suggesting that the solitary LTR has arisen relatively recently. The average heterozygosity values for each locus also varied, from 0.217 for the HERV-K113 locus to 0.00 for the monomorphic loci HERV-K107 and 3q27.2. Only one significant departure from Hardy–Weinberg equilibrium was observed in 24 individual tests; this was for the HERV-K106 solitary LTR in the Papua New Guinean population (data not shown). As 1 of 20 tests are expected to be significant at the 5% level by chance alone, this departure may be due to random statistical fluctuation. The between-population differentiation values for each bi-allelic locus ranged from 0.098 for the HERV-K115 to 0.006 for the HERV-K106 solitary LTR and were all significant by contingency analysis (data not shown). This implies that 90.2 to 99.4% of the genetic variation associated with the polymorphic HML-2 loci is within a population, supporting a recent demographic expansion of contemporary human populations.

Discussion

HERV elements make up a significant proportion of the human genome (8%) and have been proposed to be pacemakers in the evolution of primates (Sverdlov 2000). Determining the structure and cytogenetic location of HML-2 sequences that are unique to humans can be regarded as a starting point for studies investigating their impact, perhaps in regulating the expression of cellular genes or in remodeling the human genome. Here we have reported the structure and cytogenetic location of 74 human-specific HML-2 sequences, of which 15 are complete proviruses and 3 sequences represent near-complete proviral sequences which have lost one of their LTRs (Turner et al. 2001; Barbulescu et al. 1999; Costas 2001; Hughes and Coffin 2001; Sugimoto et al. 2001; Tonjes et al. 1999; Reus et al. 2001a). A single SVA retrotransposon was also characterized, which is located at Xq26.3. Intraelement homologous recombination between the 5′- and the 3′-LTRs of a provirus results in the excision of the retrovirus-like sequence and leaves behind a solitary LTR (Mager and Goodchild 1989). In this study we also describe 49 solitary LTRs, all of which are unique to humans (Mamedov et al. 2002; Medstrand and Mager 1998; Buzdin et al. 2002, 2003; Lebedev et al. 2000; Kurdyukov et al. 2001). A further six sequences have lost the 5′ or 3′ end of their LTR sequence, so we are unable to determine if they were solitary LTRs or complete proviral elements prior to sequence loss. Within this study we have not considered HML-2 sequences which subsist solely as *gag*, *pol*, or *env* genes, although the human genome is likely to contain a significant number of such sequences (Mayer et al. 1997a,b), many of which could be unique to humans.

Copy Number of HML-2 LTRs

The higher proportion of solitary LTRs within the human lineage indicates that the recombination events which led to the loss of structural genes occurred at a faster rate than the retrotransposition, integration, and fixation of novel proviral sequences within the germ line. Further implications are that the solitary LTRs are more genetically stable and/or less deleterious than a full-length provirus and that the recombination events leading to their production are occurring in quick succession after proviral integration. As three of the seven polymorphisms identified within this study (HERV-K103 HERV-K106, and HERV-K108) originate from human-specific proviruses and are generated through intraelement recombination, this observation is confirmed. Interestingly, only 1 individual possessed the HERV-K108 solitary LTR, whereas 46 individuals possessed at least one copy of the reciprocal tandem repeat, perhaps suggesting that the tandem provirus is more genetically stable than the solitary LTR. The increase in HERV-K copy number within the human lineage may also be attributed to complex and recurrent DNA arrangements such as duplication (Medstrand and Mager 1998; Nadezhdin et al. 2001) and is exemplified by the solitary LTR at 6p21.32 (Z80898), which is reported to have arisen through the duplication of the MHC complex (Horton et al. 1998).

Interlocus Recombination

In addition to operating as insertional mutagens, retroelements also serve as substrates for gene conversion and recombination, which has led to a variety of human diseases (Stankiewicz and Lupski 2002; Deininger and Batzer 2002; Ostertag et al. 2003). With the exception of a recombination event between

two HERV15 proviruses that flank the AZFa region on the human Y chromosome (Sun et al. 2000; Bosch and Jobling 2003), interelement/interlocus recombination between HERVs is not a frequent cause of human mutation. Despite this, HERV sequences are highly recombingenic (Johnson and Coffin 1999; Hughes and Coffin 2001). The recently recharacterized family of retrotransposons, SVA (SINE, VNTR, and Alu), is derived from SINE and HERV-K(HML-2) elements (Zhu et al. 1994; Ostertag et al. 2003) and a chimeric HERV-H/HERV-K retroelement transposed onto chromosomes 10, 19, and Y before the divergence of the human/chimpanzee gorilla lineages (Lapuk et al. 1999). Recombination or gene conversion has led to the concerted evolution of the HERV-H family (Mager and Freeman 1995) and has also resulted in the homogenization of the LTRs of the RTVL-1a and HERV-K(HML-2) K110/K18 proviral loci (Johnson and Coffin 1999).

Disparate target site duplications are proposed to serve as a signature of involvement of a HERV proviral sequence in interelement/interlocus recombination events. As at least 16% of the HERV-K(HML-2) proviruses that are present within the human genome are estimated to have been involved in such events, they may have had a major impact in primate genome evolution by mediating large-scale chromosomal rearrangements (Hughes and Coffin 2001). We analyzed the direct repeats of all human-specific HML-2 sequences that could be determined to have arisen through the retrotransposition of a HML-2 proviral genome, in order to assess their effect upon the plasticity of the hominid genome. Of the 63 elements that could be considered, two solitary LTRs had disparate target site duplications. PCR amplification and sequencing of their respective preintegration sites in nonhuman primates revealed that neither of these HML-2 loci had been involved in interlocus recombination events. The most parsimonious explanation for the flanking sequence duplication of the solitary LTR at 7p21.2 (Fig. 2A) and the provirus at 21q21.1 (Fig. 2B) is that unequal crossover occurred within a common ancestor who was heterozygous in possessing an allele of the preintegration site sequence and a second allele containing the integrated provirus. If such an event did occur, then the reciprocal sequence would be expected to appear as a preintegration site sequence with a deletion immediately upstream of the site of integration. In the case of the provirus at 21q21.1, in addition to a 450-bp deletion of host chromosomal DNA, the reciprocal would also contain the last 712 bp of the 3′-LTR. We also screened the human genome databases for the expected reciprocal products of unequal crossover and did not detect any such sequences, indicating either that they were not present in the representative individuals sequenced by the human genome projects or that they no longer formed a constituent of the contemporary human gene pool.

Sequence analysis of the preintegration site of the solitary LTR at 17q22 revealed that the downstream direct repeat was 4 bp shorter than the upstream (Fig. 3). The downstream target site duplication could either have lost 4 bp through deletion or during integration when an incomplete target site duplication of only 2 bp was generated. As our results show that disparate direct repeats do not always reflect interlocus recombination events and that at least 3% (2 of 63) of HERV-K(HML-2) sequences, which arose through the retrotransposition of a proviral genome, are a result of unequal crossover, the prediction that at least 16% of HML-2 proviruses have been involved in interlocus recombination events during primate genome evolution may be an overestimate. However, it should be considered that the genomic retroviral elements that exist today represent only a small fraction of total germ line integration and subsequent recombination events that have occurred, namely, those that were not detrimental to the host and that also became fixed in the genome of common ancestors.

Gene Conversion of HML-2 Proviral Loci

Mobile element families are expected to evolve following a "master gene" model of retrotransposition, whereby a few "master" elements give rise to the vast majority of novel sequences with subfamilies evolving either through the accumulation of mutations within the master genes or by the successive replacement of master genes by novel ones. Sequence exchange or gene conversion between different subfamilies of elements can confuse the expected topology, resulting in the apparent accelerated or decelerated evolution of a family (Shih et al. 1991; Mager and Freeman 1995; Kass et al. 1995; Roy-Engel et al. 2002). The phylogeny of HERV-K(HML-2) LTRs presented in this study suggests that a high degree of gene conversion has occurred within the human lineage (Fig. 4). First, the distribution of Type I proviral genomes is not monophyletic, as would be expected if these novel insertions arose from the clonal expansion of a master proviral genome which carried a 292-bp deletion at the *pol–env* boundary. We also observed similar topology for the *gag*, *pol*, and *env* structural genes (data not shown), indicating that sequence exchange was not restricted to the LTRs. This suggests that the diagnostic 292-bp deletion has been exchanged several times within recent evolutionary time scale between the Type I and the Type II genomes, leading to the production of mosaic proviruses (Costas 2001). Second, the classification of HML-2 LTRs into the subtypes Hs-a and Hs-b (Buzdin et al. 2003) is also not consistent with clonal expansion, as the subtypes group independently of the Type I and

Type II proviral genomes. During sequence analysis we also observed that none of the human-specific LTRs contained the diagnostic 8- and 23-bp insertions that are present within the LTRs of the HML-2 ancestor sequences, HERV-K(OLD) (Reus et al. 2001b), indicative of sequence exchange exclusively between highly homologous and recently retrotransposed sequences. The divergent LTRs of the HERV-K115 provirus support this view, as the element is predicted to have entered the human gene pool relatively recently (Turner et al. 2001). However, if gene conversion has occurred as frequently as the Type I/Type II phylogeny suggests, then additional HERV-K(HML-2) proviruses would also be expected to possess highly divergent LTRs. Gene conversion leading to the homogenization of the LTRs within a provirus would counteract this effect and is likely to have occurred regularly within the human lineage as exemplified by the provirus at 12q14.1. This phenomenon has previously been observed within the HERV-K(HML-2) K110/K18 and RTVL-1a proviral loci (Johnson and Coffin 1999). A major consequence of such gene conversion events would be that the accumulative nucleotide differences between the two LTRs of a provirus do not accurately reflect the relative age of the provirus, as demonstrated by the provirus at 11q23.2, which represents an underestimate of time since integration.

HERV-K(HML-2) Polymorphisms for the Study of Human Evolution

HERV insertional or structural mutations leading to the production of a solitary LTR offer several advantages for examining human genomic diversity. First, large numbers of DNA samples can be rapidly typed using PCR-based assays. Second, as with LINE and SINE retroelements, the *de novo* insertion of a HERV sequence within the germ line represents a unique event in human genome evolution. The large number of potential target sites within the human genome and the random nature of retroviral integration denote that homoplasy is highly unlikely. Third, HERV sequences are stable, as there are no known mechanisms for completely removing them without deleting host chromosomal DNA or leaving behind a solitary LTR. Accordingly, the directionality of the insertion and the formation of a solitary LTR can unambiguously be assigned to a specific lineage, as individual loci containing the same HERV sequence are identical by descent. Fourth, the ancestral state of a HERV sequence is ultimately its absence and is represented by a preintegration site sequence. HERV sequences that are unique to humans can be determined through PCR analysis of the orthologous region in nonhuman primates. This information can be used to root trees of population relationships derived from analysis of HERV polymorphisms. Finally, as the process of reverse transcription generates two LTR sequences that are identical at the time of HERV sequence insertion, the accumulative nucleotide differences between them can serve as a molecular clock (Dangel et al. 1995). However, this measure will be invalidated if a HERV sequence has been subject to recombination or gene conversion after integration.

To ascertain the utility of HERV polymorphisms for examining human evolution, we screened each of the 74 human-specific sequences reported within this study for polymorphism within the human genome databases and determined that seven HML-2 elements were dimorphic. Two of these were solitary LTRs which were polymorphic for insertion. The first is located at 6p21.32 and is reported to have arisen through the duplication of the MHC complex (Horton et al. 1998); the second is located at 9q12. As this chromosomal region is highly repetitive, it is impossible to confirm through PCR amplification the allelic variation of the loci.

In order to examine the genomic variation associated with the remaining five polymorphic HML-2 loci, we developed a PCR-based assay to determine the allelic variation associated with each of them in four diverse human populations. Three bi-allelic loci, HERV-K113, HERV-K115, and HERV-K108, showed geographical distributions that were consistent with previous reports (Reus et al. 2001a; Turner et al. 2001). The two remaining loci, HERV-K103 and HERV-K106, were dimorphic for a solitary LTR and complete copy of the provirus. The HERV-K106 solitary LTR had an average allele frequency of 0.079 and was present in all populations, whereas the HERV-K103 solitary LTR was only present in the heterozygous state in two African individuals. This indicates that the HERV-K103 solitary LTR may be a structural mutation that has arisen relatively recently or that it was unfixed at the time of human expansion from Africa.

Computational screening for the detection of novel retroelement insertion has previously been observed to be subject to bias (Myers et al. 2002). This dictates that high-frequency polymorphisms are lost in the screening process and low-frequency polymorphisms are underrepresented in the human genome databases. We surveyed the genomic variation associated with a further two human-specific HML-2 proviral loci, HERV-K107 and 3q27.2, in order to assess if polymorphism was present. According to the accumulative nucleotide divergence of their respective LTRs, both of these proviruses entered the human gene pool recently and so were likely candidates for insertional or solitary LTR allelic variation. Our results indicated that each of the loci was monomorphic, in accordance with the human genome databases. These results

further emphasize that for recent evolutionary events, the accumulative nucleotide differences of the LTRs of a HML-2 provirus do not serve as an accurate measure of time since insertion.

The debate over recent human origins has focused on two models (reviewed by Stringer 2002). The "multiregional model" proposes that over the last 1.5 million years, modern humans arose independently in different regions of the world but remained a single species through worldwide gene flow. In contrast, the "recent replacement model," or "Out of Africa 2," suggests that a single population of modern humans migrated from Africa approximately 100,000 to 200,000 years ago and replaced archaic human populations throughout the world. Our survey of the human genomic diversity of HML-2 loci indicates that the genetic diversity of the African population is far higher than non-African populations and that 90.2 to 99.4% of this genetic variability is within a population, supporting a recent demographic expansion of modern humans from Africa.

Acknowledgment. We would like to thank Alastair Macdonald, Rochelle Bleeker, Anna Meredith, Benjamin Searle, and Sonia Lee for their help with sample collection.

References

Akopov SB, Nikolaev LG, Khil PP, Lebedev YB, Sverdlov ED (1998) Long terminal repeats of human endogenous retrovirus K family (HERV-K) specifically bind host cell nuclear proteins. FEES Lett 421:229–233

Barbulescu M, Turner G, Seaman MI, Deinard AS, Kidd KK, Lenz J (1999) Many human endogenous retrovirus K (HERV-K) proviruses are unique to humans. Curr Biol 9:861–868

Bosch E, Jobling MA (2003) Duplications of the AZFa region of the human Y chromosome are mediated by homologous recombination between HERVs and are compatible with male fertility. Hum Mol Genet 12:341–347

Buzdin A, Khodosevich K, Mamedov I, Vinogradova T, Lebedev Y, Hunsmann G, Sverdlov E (2002) A technique for genome-wide identification of differences in the interspersed repeats integrations between closely related genomes and its application to detection of human-specific integrations of HERV-K LTRs. Genomics 79:413–422

Buzdin A, Ustyugova S, Khodosevich K, Mamedov I, Lebedev Y, Hunsmann G, Sverdlov E (2003) Human-specific subfamilies of HERV-K (HML-2) long terminal repeats: three master genes were active simultaneously during branching of hominoid lineages. Genomics 81:149–156

Chen FC, Li WH (2001) Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. Am J Hum Genet 68:444–456

Costas J (2001) Evolutionary dynamics of the human endogenous retro virus family HERV-K inferred from full-length proviral genomes. J Mol Evol 53:237–243

Dangel AW, Baker BJ, Mendoza AR, Yu CY (1995) Complement component C4 gene intron 9 as a phylogenetic marker for primates: Long terminal repeats of the endogenous retrovirus ERV-K(C4) are a molecular clock of evolution. Immunogenetics 42:41–52

Deininger PL, Batzer MA (2002) Mammalian retroelements. Genome Res 12:1455–1465

Domansky AN, Kopantzev EP, Snezhkov EV, Lebedev YB, Leib-Mosch C, Sverdlov ED (2000) Solitary HERV-K LTRs possess bi-directional promoter activity and contain a negative regulatory element in the U5 region. FEES Lett 472:191–195

Faerman M, Filon D, Kahila G, Greenblatt CL, Smith P, Oppenheim A (1995) Sex identification of archaeological human remains based on amplification of the X and Y amelogenin alleles. Gene 167:327–332

Frazer KA, Chen X, Hinds DA, Pant PV, Patil N, Cox DR (2003) Genomic DNA insertions and deletions occur frequently between humans and nonhuman primates. Genome Res 13:341–346

Goodchild NL, Freeman JD, Mager DL (1995) Spliced HERV-H endogenous retroviral sequences in human genomic DNA: Evidence for amplification via retrotransposition. Virology 206:164–173

Horton R, Niblett D, Milne S, Palmer S, Tubby B, Trowsdale J, Beck S (1998) Large-scale sequence comparisons reveal unusually high levels of variation in the HLA-DQB1 locus in the class II region of the human MHC. J Mol Biol 282:71–97

Hughes JF, Coffin JM (2001) Evidence for genomic rearrangements mediated by human endogenous retroviruses during primate evolution. Nat Genet 29:487–489

Huh JW, Hong KW, Yi JM, Kirn TH, Takenaka O, Lee WH, Kim HS (2003) Molecular phylogeny and evolution of the human endogenous retrovirus HERV-W LTR family in hominoid primates. Mol Cells 15:122–126

International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. Nature 409:860–921

Johnson WE, Coffin JM (1999) Constructing primate phylogenies from ancient retrovirus sequences. Proc Natl Acad Sci USA 96:10254–10260

Kass DH, Batzer MA, Deninger PL (1995) Gene conversion as a secondary mechanism in SINE evolution. Mol Cell Biol 15:19–25

Kurdyukov SG, Lebedev YB, Artamonova II, Gorodentseva TN, Batrak AV, Mamedov IZ, Azhikina TL, Legchilina SP, Efimenko IG, Gardiner K, Sverdlov ED (2001) Full-sized HERV-K (HML-2) human endogenous retroviral LTR sequences on human chromosome 21: map locations and evolutionary history. Gene 273:51–61

Lapuk AV, Khil PP, Lavrentieva IV, Lebedev YB, Sverdlov ED (1999) A human endogenous retrovirus-like (HERV) LTR formed more than 10 million years ago due to an insertion of HERV-H LTR into the 5′ LTR of HERV-K is situated on human chromosomes 10, 19 and Y. J Gen Virol 80:835–839

Lavrentieva I, Khil P, Vinogradova T, Akhmedov A, Lapuk A, Shakhova O, Lebedev Y, Monastyrskaya G, Sverdlov ED (1998) Subfamilies and nearest-neighbour dendrogram for the LTRs of human endogenous retroviruses HERV-K mapped on human chromosome 19: Physical neighbourhood does not correlate with identity level. Hum Genet 102:107–116

Lebedev YB, Belonovitch OS, Zybrova NV, Khil PP, Kurdyukov SG, Vinogradova TV, Hunsmann G, Sverdlov ED (2000) Differences in HERV-K LTR insertions in orthologous loci of humans and great apes. Gene 247:265–277

Liao D, Pavelitz T, Weiner AM (1998) Characterization of a novel class of interspersed LTR elements in primate genomes: Structure, genomic distribution, and evolution. J Mol Evol 46:649–660

Liu G, Zhao S, Bailey JA, Sahinalp SC, Alkan C, Tuzun E, Green ED, Eichler EE (2003) Analysis of primate genomic variation

reveals a repeat-driven expansion of the human genome. Genome Res 13:358–368

Locke DP, Segraves R, Carbone L, Archidiacono N, Albertson DG, Pinkel D, Eichler EE (2003) Large-scale variation among human and great ape genomes determined by array comparative genomic hybridization. Genome Res 13:347–357

Lower R, Lower J, Kurth R (1996) The viruses in all of us: Characteristics and biological significance of human endogenous retrovirus sequences. Proc Natl Acad Sci USA 93:5177–5184

Mager DL, Freeman DJ (1995) HERV-H endogenous retroviruses: Presence in the new world branch but amplification in the old world primate lineage. Virology 213:395–404

Mager DL, Goodchild NL (1989) Homologous recombination between the LTRs of a human retrovirus-like element causes a 5-kb deletion in two siblings. Am J Hum Genet 45:848–854

Mamedov I, Batrak A, Buzdin A, Arzumanyan E, Lebedev Y, Sverdlov ED (2002) Genome-wide comparison of differences in the integration sites of interspersed repeats between closely related genomes. Nucleic Acids Res 30:e71

Mayer J, Meese E, Mueller-Lantzsch N (1997a) Chromosomal assignment of human endogenous retrovirus K (HERV-K) env open reading frames. Cytogenet Cell Genet 79:157–161

Mayer J, Meese E, Mueller-Lantzsch N (1997b) Multiple human endogenous retrovirus (HERV-K) loci with gag open reading frames in the human genome. Cytogenet Cell Genet 78:1–5

Mayer J, Meese E, Mueller-Lantzsch N (1998) Human endogenous retrovirus K homologous sequences and their coding capacity in Old World primates. J Virol 72:1870–1875

Medstrand P, Blomberg J (1993) Characterization of novel reverse transcriptase encoding human endogenous retroviral sequences similar to type A and type B retroviruses: Differential transcription in normal human tissues. J Virol 67:6778–6787

Medstrand P, Mager DL (1998) Human-specific integrations of the HERV-K endogenous retrovirus family. J Virol 72:9782–9787

Myers JS, Vincent BJ, Udall H, Watkins WS, Morrish TA, Kilroy GE, Swergold GD, Henke J, Henke L, Moran JV, Jorde LB, Batzer MA (2002) A comprehensive analysis of recently integrated human Ta L1 elements. Am J Hum Genet 71:312–326

Nadezhdin EV, Lebedev YB, Glazkova DV, Bornholdt D, Arman IP, Grzeschik KH, Hunsmann G, Sverdlov ED (2001) Identification of paralogous HERV-K LTRs on human chromosomes 3, 4, 7 and 11 in regions containing clusters of olfactory receptor genes. Mol Genet Genomics 265:820–825

Ono M, Yasunaga T, Miyata T, Ushikubo H (1986) Nucleotide sequence of human endogenous retrovirus genome related to the mouse mammary tumor virus genome. J Virol 60:589–598

Ostertag EM, Goodier JL, Zhang Y, Kazazian HH (2003) SVA elements are nonautonomous retrotransposons that cause disease in humans. Am J Hum Genet 73:1444–1451

Patience C, Wilkinson DA, Weiss RA (1997) Our retroviral heritage. Trends Genet 13:116–120

Reus K, Mayer J, Sauter M, Scherer D, Muller-Lantzsch N, Meese E (200la) Genomic organization of the human endogenous retrovirus HERV-K(HML-2.HOM) (ERVK6) on chromosome 7. Genomics 72:314–320

Reus K, Mayer J, Sauter M, Zischler H, Muller-Lantzsch N, Meese E (2001b) HERV-K (OLD): Ancestor sequences of the human endogenous retrovirus family HERV-K (HML-2). J Virol 75:8917–8926

Roy-Engel AM, Carroll ML, El-Sawy M, Salem A, Garger RK, Nguyen SV, Deininger PL, Batzer MA (2002) Non-traditional *Alu* evolution and primate genomic diversity. J Mol Biol 316:1033–1040

Seifarth W, Baust C, Murr A, Skladny H, Krieg-Schneider F, Blusch J, Werner T, Hehlmann R, Leib-Mosch C (1998) Proviral structure, chromosomal location, and expression of HERV-K- T47D, a novel human endogenous retrovirus derived from T47D particles. J Virol 72:8384–8391

Shen L, Wu LC, Sanlioglu S, Chen R, Mendoza AR, Dangel AW, Carroll MC, Zipf WB, Yu CY (1994) Structure and genetics of the partially duplicated gene RP located immediately upstream of the complement C4A and C4B genes in the HLA class III region: Molecular cloning, exon-intron structure, composite retroposon, and breakpoint of gene duplication. J Biol Chem 269:8466–8476

Shih A, Coutavas EE, Rush MG (1991) Evolutionary implications of primate endogenous retroviruses. Virology 185:495–502

Simmonds P, Smith DB (1999) Structural constraints on RNA virus evolution. J Virol 73:5787–5794

Simpson GR, Patience C, Lower R, Tonjes RR, Moore HD, Weiss RA, Boyd MT (1996) Endogenous D-type (HERV-K) related sequences are packaged into retroviral particles in the placenta and possess open reading frames for reverse transcriptase. Virology 222:451–456

Stankiewicz P, Lupski JR (2002) Molecular-evolutionary mechanisms for genomic disorders. Curr Opin Genet Dev 12:312–319

Stringer C (2002) Modern human origins: progress and prospects. Philos Trans R Soc Lond B Biol Sci 357:563–579

Sugimoto J, Matsuura N, Kinjo Y, Takasu N, Oda T, Jinno Y (2001) Transcriptionally active HERV-K genes: Identification, isolation, and chromosomal mapping. Genomics 72:137–144

Sun C, Skaletsky H, Rozen S, Gromoll J, Nieschlag E, Oates R, Page DC (2000) Deletion of azoospermia factor a (AZFa) region of human Y chromosome caused by recombination between HERV15 proviruses. Hum Mol Genet 9:2291–2296

Sverdlov ED (2000) Retroviruses and primate evolution. Bioessays 22:161–171

Tonjes RR, Czauderna F, Kurth R (1999) Genome-wide screening, cloning, chromosomal assignment, and expression of full-length human endogenous retrovirus type K. J Virol 73:9187–9195

Towler EM, Gulnik SV, Bhat TN, Xie D, Gustschina E, Sumpter TR, Robertson N, Jones C, Sauter M, Mueller-Lantzsch N, Debouck C, Erickson JW (1998) Functional characterization of the protease of human endogenous retrovirus, K10: Can it complement HIV-1 protease? Biochemistry 37:17137–17144

Tristem M (2000) Identification and characterization of novel human endogenous retrovirus families by phylogenetic screening of the human genome mapping project database. J Virol 74:3715–3730

Turner G, Barbulescu M, Su M, Jensen-Seaman MI, Kidd KK, Lenz J (2001) Insertional polymorphisms of full-length endogenous retroviruses in humans. Curr Biol 11:1531–1535

Urnovitz HB, Murphy WH (1996) Human endogenous retroviruses: nature, occurrence, and clinical implications in human disease. Clin Microbiol Rev 9:72–99

Vinogradova TV, Leppik LP, Nikolaev LG, Akopov SB, Kleiman AM, Senyuta NB, Sverdlov ED (2001) Solitary human endogenous retroviruses-K LTRs retain transcriptional activity in vivo, the mode of which is different in different cell types. Virology 290:83–90

Zhu ZB, Jian B, Volanakis JE (1994) Ancestry of SINE-R.C2 a human-specific retroposon. Hum Genet 93:545–551

Zsiros J, Jebbink MF, Lukashov VV, Voute PA, Berkhout B (1998) Evolutionary relationships within a subgroup of HERV-K-related human endogenous retroviruses. J Gen Virol 79:61–70

Zsiros J, Jebbink MF, Lukashov VV, Voute PA, Berkhout B (1999) Biased nucleotide composition of the genome of HERV-K related endogenous retroviruses and its evolutionary implications. J Mol Evol 48:102–111